UNRAVELING THE HISTORY OF DEFORESTATION IN THE AMAZON RAINFOREST WITH STATISTICAL MODELING

A Thesis

presented to

the Faculty of California Polytechnic State University,

San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Statistics

by

Ryan DeStefano

June 2024

COMMITTEE MEMBERSHIP

TITLE: Unraveling the History of Deforestation in the Amazon Rainforest with Statistical Modeling

AUTHOR: Ryan DeStefano

DATE SUBMITTED: June 2024

COMMITTEE CHAIR: Hunter Glanz, Ph.D.
Professor of Statistics and Data Science

COMMITTEE MEMBER: Andrew Fricker, Ph.D.
Professor of Geography

COMMITTEE MEMBER: Soma Roy, Ph.D.
Professor of Statistics

ABSTRACT

Unraveling the History of Deforestation in the Amazon

Rainforest with Statistical Modeling

Ryan DeStefano

The Amazon rainforest, a vital ecosystem of immense biodiversity and global climate significance, faces the ongoing threat of deforestation driven by agricultural expansion. This thesis employs remote sensing techniques, focusing on the Enhanced Vegetation Index (EVI) derived from Landsat satellite imagery, to track land cover dynamics within the Amazon. The study examines historical land cover changes in current plantations in Peru and Brazil, regions where the exact timing of deforestation is uncertain. By analyzing EVI measurements dating back to 1984, inflection points indicative of deforestation events preceding plantation establishment are identified. Statistical modeling techniques, including spline fitting to analyze time series data and Random Forest algorithms for calibration, are employed to enhance the accuracy of EVI measurements. Additionally, predictions for deforestation years derived from ALOS satellite data are compared with those from Landsat imagery, revealing discrepancies and underscoring the need for methodological refinement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

The Amazon rainforest stands as a critical ecosystem, harboring rich biodiversity and playing a significant role in global climate regulation. However, the persistent threat of deforestation, primarily driven by agricultural expansion, poses a risk to this important area. Understanding the dynamics of land cover change within the Amazon is thus of paramount importance for effective conservation and sustainable land management efforts as well as understanding the patterns and seasonality of land cover growth.

In recent years, advancements in remote sensing technology have revolutionized the ability to monitor and analyze changes in land cover at a large scale. Leveraging satellite imagery with tools such as Google Earth Engine (GEE), researchers now have access to vast amounts of data and can now obtain valuable insights into the spatial and temporal patterns of land use and land cover change with high accuracy, being able to extract a multitude of metrics from various satellites with relative ease.

This thesis focuses on the application of remote sensing techniques, particularly the measurement of the Enhanced Vegetation Index (EVI), to track land cover dynamics within the Amazon rainforest. EVI, derived from the Landsat satellite series, serves as a reliable indicator of vegetation density and health, providing information on the extent and intensity of forest cover.

The primary objective of this research is to investigate historical land cover changes in targeted areas within the Amazon, particularly in regions where extensive deforestation has occurred to make way for agricultural plantations. Of particular interest are thousands of plantations scattered across Brazil and Peru, where the historical timeline of deforestation and plantation establishment remains largely uncharted.

By delving into the historical EVI measurements dating back to 1984, this study aims to identify crucial inflection points indicative of deforestation activities preceding the establishment of plantations. The detection of minimum EVI values serves as a barometer for pinpointing the timing of deforestation events, enabling the assessment of when these areas were cleared. This process is very similar to that conducted by Logan Berner in his paper "LandsatTS: an R package to facilitate retrieval, cleaning, cross-calibration, and phenological modeling of Landsat time series data". Berner's paper is focused on analysis of a boreal forest and thus uses different methodology than is what is required for a rich and dense rainforest such as the Amazon. A goal of this paper is to adapt his methods to be geared towards analysis of deforestation in dense rainforest climates.

To address challenges such as cloud cover and other atmospheric conditions that may obscure satellite imagery, statistical modeling techniques will be employed to fit time series data and fill in data gaps. Moreover, measurements from the Advanced Land Observing Satellite, which captures data in both horizontal transmit and horizontal receive (HH) and horizontal transmit and vertical receive (HV) polarizations, will be utilized as a supplementary analysis to give an alternative lens to the history of the areas of interest.

Through a multidisciplinary approach combining remote sensing, statistical modeling, and geographic information systems (GIS), this study seeks to uncover the motivations behind land cover changes in the Amazon rainforest. By illuminating the connections among deforestation, the establishment of plantations, and environmental processes, it is possible to gain a large amount of information on the history of this ecosystem.

Chapter 2 of this thesis gives a detailed overview of the machinery behind remote sensing as well as many extended definitions of terms mentioned throughout the paper. Chapter 3 details a comprehensive literature review of related works in remote sensing. Chapter 4 discusses the details/steps of the analysis. Chapter 5 highlights the main results of the analysis. Chapter 6 concludes with closing remarks, limitations, and directions for future work.

Chapter 2

BACKGROUND

This chapter aims to define important terminology associated with the research that will be used throughout the paper, as well as contextualize how these terms/topics integrate with the broader research goals.

2.1 Landsat Satellites

The Landsat program consists of a series of Earth-observing satellite missions jointly managed by NASA and the U.S. Geological Survey [1]. There have been 10 Landsat satellites deployed since 1972, all of which contain slight improvements over the previous versions and are named successively. For the purposes of our research, we obtained measurements from Landsat 5, Landsat 7, and Landsat 8. It is our understanding that all the plantations in this study were deforested from 1984 on, which is why we started with measurements from Landsat 5 which started around that time. The leapfrog over Landsat 6 was due to this satellite being unsuccessfully launched into orbit [1]. Each of these satellites contains a series of spectral bands, sensors within the satellite that receive energy from sun reflectance [2]. Each of these bands picks up different wavelengths and thus measure different things. Put together, the bands can represent different metrics within the remote sensing sphere. The metric of interest for this research specifically is Enhanced Vegetation Index (EVI).

2.1.1 Enhanced Vegetation Index

EVI is a measurement within the remote sensing world used to quantify the greenness of vegetation and is the variable driving the conclusions of this research. This metric is derived from a more commonly used value, Normalized Difference Index, but is optimal for this issue because of the nature of the Amazon. EVI corrects for some atmospheric conditions and canopy background noise and is more sensitive in areas with dense vegetation [15]. Thus, EVI is well suited to use as the metric of interest for the Amazon which satisfies these exact conditions, having extremely dense and rich vegetation. EVI is calculated using the Red, Blue, and NIR bands from the Landsat satellites, with formula specified below.

$$2.5 * \frac{NIR - Red}{(NIR + 6)(Red - 7.5)(Blue + 1)}$$

Large Enhanced Vegetation Index values correspond to areas with more photosynthetic activity and dense land cover [3].

2.2 Advanced Land Observing Satellite

The Advanced Land Observing Satellite (ALOS) program comprises a constellation of Earth observation satellites developed by the Japan Aerospace Exploration Agency. The ALOS program aims to facilitate monitoring of Earth's surface and address various environmental and societal challenges [4]. The ALOS series includes ALOS-1, which was equipped with three remote sensing instruments: the Advanced Visible and Near

Infrared Radiometer type-2, the Panchromatic Remote-sensing Instrument for Stereo Mapping, and the Phased Array type L-band Synthetic Aperture Radar (PALSAR). These instruments provided high-resolution optical and radar imagery, enabling detailed observations of land cover, terrain, and environmental changes. A successor mission, ALOS-2, was launched in 2014, featuring an upgraded L-band Synthetic Aperture Radar (SAR) instrument known as PALSAR-2. The latter, ALOS-2 PALSAR-2 is the satellite used in this research to extract measurements and aid in giving an alternative view to Landsat on the land cover phenomena in the Amazon. The ALOS-2 satellite has bands different from that of Landsat, and of which we are interested are bands Horizontal Transmit-Horizontal Receive (HH) and Horizontal Transmit-Vertical Receive (HV). These HH and HV polarizations are measured through radar signals being shot to the surface and bouncing back to the satellite. The use of the measurements from the ALOS satellite allows an alternative view on the land cover of the plantations as ALOS utilizes microwave radar to obtain measurements.

2.2.1 Horizontal-Horizontal Polarization

In the HH polarization mode, the ALOS radar signal is transmitted in a horizontal orientation and received back in a horizontal orientation. This configuration is sensitive to surface roughness and scattering properties of vegetation and bare soil. Essentially, this band is measuring the roughness of the ground level surface and canopy. A large value corresponds to high surface roughness while a low value corresponds to low surface roughness. HH also provides information about moisture content, which is also important to account for when assessing land cover [4]. Wetter areas tend to have lower HH values

because the wetness absorbs some of the radar energy. It is important to consider this aspect along with the surface roughness when assessing any HH measurement.

2.2.2 Horizontal-Vertical Polarization

In the HV polarization mode, the radar signal is transmitted in a horizontal orientation but received back in a vertical orientation. This polarization is sensitive to volume scattering, particularly from vegetation canopy layers. HV polarization can help differentiate between different types of vegetation and assess their biomass and structural properties, a main aspect being able to assess the height of vegetation in any given area [4]. An intuitive way to think of HV measurements is that areas with tall trees will return large HV values while deforested areas or those with short vegetation will return small values.

2.3 Google Earth Engine

Google Earth Engine is the source used for extracting data for the areas of interest within the Amazon. GEE contains a vast amount of geospatial data from different satellites, including Landsat and ALOS [16]. The Google Earth Engine API was called from within R, leveraging the rgee package to interact with GEE, to obtain all measurements in this research [17].

Chapter 3

LITERATURE REVIEW

The subsequent literature review delves into previous research relevant to the study at hand, aiming to explain the methodologies, findings, and insights obtained from similar research. By examining a variety of research papers, noteworthy outcomes contributing to the broader research context will be featured. This review will be organized into subsections, each focusing on specific themes found in the papers, synthesizing the results.

3.1 Agricultural Practices

In the assessment of agricultural practices in tropical regions, researchers have focused on understanding the dynamics of key crops such as palm oil and cocoa, native vegetation to tropical regions which are crucial for local economies and have significant environmental implications [6]. Khiabani et al. conducted an extensive evaluation of palm oil yield and biophysical suitability in Indonesia and Malaysia. Their study provided understanding into the factors influencing palm oil yield variations across different regions, underlining the importance of soil properties, climate conditions, and management practices [5]. Through remote sensing data analysis and computational modeling, they uncovered the disparities between actual and potential yields, with

Malaysian states generally achieving higher actual yields compared to Indonesian provinces.

Similarly, Vera-Velez et al. investigated cocoa agroforestry systems in the Amazon region, aiming to understand their role in mitigating forest degradation resulting from shifting agriculture practices [6]. They demonstrated that agroforestry systems not only maintain higher rates of native and endangered species but also contribute to forest conservation efforts. By preserving arboreal structures and promoting biodiversity, these systems serve as alternatives to conventional agricultural practices, emphasizing the importance of integrating environmental conservation with agricultural production [6]. Furthermore, Berner et al. noted the potential of remote sensing technologies in monitoring land use changes associated with cocoa agroforestry systems, explaining the spatial distribution and dynamics of land cover types within agricultural landscapes. Through the synthesis of findings from these studies, it becomes evident that sustainable agricultural practices in tropical regions play a critical role in enhancing productivity, preserving biodiversity, and mitigating environmental degradation.

3.2 Land Cover Mapping

Remote sensing techniques have revolutionized land cover mapping and monitoring, offering a look into landscape dynamics and ecosystem health. Pereira et al. evaluated the performance of ALOS/PALSAR data for land cover mapping in the Amazon, revealing its potential for large-area classification and accurate forest cover estimation [7]. Leveraging radar sensors, their study overcame limitations associated with cloud cover, stemming from the use of SAR radar rather than Landsat. Berner et al. developed the

"LandsatTS" R package to facilitate the retrieval, cleaning, and analysis of Landsat time series data, enabling broader use of Landsat satellite data for assessing ecosystem history over the past four decades [9]. This technique and package are drawn off heavily for the research discussed in this paper.

By integrating remote sensing data with computational tools, researchers can enhance the accuracy of land cover mapping, providing information for a variety of uses, including environmental management and conservation efforts. Walker et al. explained the importance of multi-sensor data fusion for comprehensive land cover characterization, emphasizing the complementary nature of radar and optical sensors in capturing spatial and temporal variations in land cover types, that is, comparing different spectral bands within a satellite [8]. Through the synthesis of these findings, it is evident that remote sensing technologies, and the fusion of multiple spectral bands within a satellite, have the ability to play a crucial role in monitoring land cover changes in a variety of regions and ecosystems.

3.3 Comparing Remote Sensing Data Sources

Comparative analyses of remote sensing data sources have proven to be an informative strategy for land cover monitoring applications. Pereira et al. compared Landsat and ALOS data for forest cover mapping in the Brazilian Amazon, highlighting the complementary nature of radar and optical sensors, comparing Landsat to ALOS results. While Landsat data excelled in discriminating densely forested classes, radar data performed better for non-densely forested classes, such as pastures and agricultural lands [7]. Similarly, Walker et al. emphasized the importance of integrating multi-sensor and

multi-temporal data for more accurate forest monitoring systems in tropical regions, specifically due to the rich nature of vegetation in these types of areas [8].

By leveraging the strengths of different remote sensing platforms, it is possible to obtain a clearer view of the inner workings of ecosystems. These studies underscore the importance of integrating radar and optical data sources to overcome inherent limitations associated with individual sensor platforms, ultimately contributing to improved understanding and management of ecosystems.

3.4 Advancements in Remote Sensing

Advancements in remote sensing technologies have expanded the capabilities of satellite sensors for land cover monitoring applications. Walker et al. discussed the increasing accessibility and utility of synthetic aperture radar (SAR) data for assessments of forest cover and ecosystem function and emphasized the importance of long-term data records from radar sensors for building reliable forest monitoring systems [8]. And emphasizing the potential of SAR data to complement optical sensors in assessing ecosystem dynamics.

These advancements underscore the growing role of remote sensing in providing a clearer view of the overall landscape of the Earth. As technology continues to evolve, remote sensing is poised to play an increasingly significant role in monitoring and managing Earth's ecosystems, ultimately contributing to informed decision-making and sustainable development practices worldwide. With the capabilities of SAR data and integration with optical and other remote sensing datasets, researchers can gather insights for addressing

global environmental challenges and promoting sustainable development. It is evident that remote sensing infrastructure will continue to advance in the coming years, leading to a better understanding of all the parameters within an area of interest. An upcoming satellite, coined NISAR, is set to release in late 2024 and is a more advanced version of ALOS which will allow for a higher resolution view of forested ecosystems [18].

Chapter 4

IMPLEMENTATION

This section primarily outlines the methods used for data collection, processing, and analysis. It will begin with a discussion of the plantations of interest; how they were chosen and some characteristics of the areas. This will be followed by a discussion of the steps for analyzing the deforestation trends in the plantations using the Landsat satellites. The section will conclude with a discussion of the steps for analyzing the data using the ALOS satellite.

4.1 Plantations of Interest

The plantations chosen for this analysis were those delineated by the scientists within NASA-SERVIR, a joint initiative between NASA and the United States Agency for International Development that leverages satellite data and Earth observation technology to aid developing countries in environmental decision-making and climate resilience, focusing on capacity building, tailored geospatial services, and regional partnerships to enhance natural resource management and disaster response [19]. These plantations are in Peru and Brazil, within areas of the Amazon that have been developed more recently, 1984 and later, and do not have succinct, accurate records of the activity, such as deforestation, that have occurred within the past 40 years.

4.1.1 Peru Plantations

There are 1,281 plantations of interest in Peru that are located around the southwest border of the Amazon. The general location of the plantations is a latitude of -8.18 and longitude of -74.95, as marked in Figure 4.1.

Figure 4.1: Peru Plantations General Location

All plantations of interest in Peru are primarily palm oil plantations, and as discussed previously, palm oil harvesting is one of the main economic factors driving deforestation. Figure 4.2 displays a portion of the 1,281 plantations and is zoomed in onto the marked location in Figure 4.1. Each of the polygons represents one of the plantations that we are interested in making deforestation predictions for. Note that there are more plantations not displayed within the Figure but that are still analyzed. The plantations differ greatly in size, the biggest being about 100 times bigger than the smallest.

Figure 4.2 Delineated Peru Plantations

4.1.2 Brazil Plantations

There are 948 plantations of interest in Brazil that are located around the eastern border of the Amazon. The general location of the plantations is a latitude of -2.77 and longitude of -51.63 as marked in Figure 4.3.

Figure 4.3: Brazil Plantations General Location

The plantations in Brazil have a variety of primary vegetation, including cocoa, palm oil, and açaí. Figure 4.4 displays a portion of the 948 plantations and is zoomed in onto the marked location in Figure 4.3. Each of the individual polygons represents a plantation of interest. Note that the Figure does not display all the plantations that were analyzed. The plantations in Brazil do differ in size but not nearly as much as in Peru.

Figure 4.4 Delineated Brazil Plantations

4.2 Landsat

This section aims to explain the steps taken to analyze deforestation trends in the plantations of interest, explaining the workflow for the Landsat analysis. Again, the main goal for the Landsat analysis is to find the year that each of the plantations was deforested, signaled by the year where the smallest Enhanced Vegetation Index was measured/predicted. Other goals include being able assess overall trends in Enhanced Vegetation Index (EVI) over the 40-year period, 1984 to 2024.

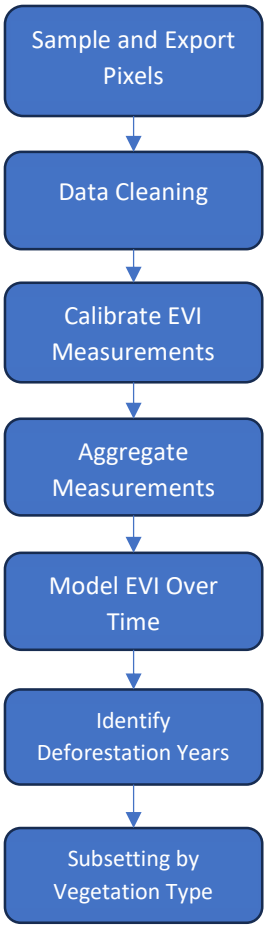


Figure 4.5 Landsat Workflow

4.2.1 LandsatTS Package

The R package, "LandsatTS", is used liberally throughout this analysis to aid in retrieving, cleaning, and modeling the EVI time series. Listed in Table 4.1 are the functions used within the package with a brief description of the functionality. More details in later sections will come on the usage and specificities of each function. Note that the functions were used in the order they are listed in the table.

Table 4.1 LandsatTS Functions

Function	Description
last_export_ts()	Export surface reflectance measurements for a set of latitude longitude pairs.
lsat_format_data()	Takes a dataframe of exported pixels and formats column names and scales values for subsequent use.
lsat_clean_data()	Filters out certain measurements due to weather factors, such as heavy rain or cloud cover.
lsat_calc_spectral_index()	Calculates specified spectral index, in our case EVI, based on spectral band measurements.
lsat_calibrate_rf()	Calibrates the differences in spectral index measurements from the different Landsat satellites using a random forest model.
lsat_fit_phenological_curves()	Fits splines to the measurements, plotting the trends in each sample point by year.

4.2.2 Sample Point Selection and Export

At this point in the process, the plantations of interest were specified and the machinery, R and the "LandsatTS" package, were in place. The next step was to get the latitude longitude pairs, essentially points/pixels, within each plantation measured by the Landsat satellites exported from Google Earth Engine using the lsat_export_ts() function within R. This function takes latitude longitude pairs along with an identifier and sends them to Google Earth Engine which then pulls each measurement from the Landsat 5, 7, and 8 satellites dating back to 1984 for each of the points. That is, for every sample point, there was a large number of measurements pulled; these measurements being the spectral band values for each point in time, that will later be used to calculate EVI. The original plan was to export every point within each plantation, but upon testing it was discovered that this was not at all feasible due to our computational limits. It took approximately one minute to fully export the measurements dating back to 1984 for two pixels, that is a rate of 2 pixels/min. We were working with over 2,000 plantations, with the number of pixels within each plantation ranging from 10 to 32,000. If every point from every plantation was exported it would take months to finish running. Thus, a system had to be put in place to decide how many points to sample from each of the plantations.

The decision of how many points to sample depended on a couple factors; the length of time to export the pixels and getting a representative sample of each plantation, the latter being more important. This involved first getting an estimate of the number of points/pixels within each of the plantations based on area within the region. Upon testing results on a series of plantations it was found that we got the same results using 10% of the pixels in each plantation as we did when using 50% of the pixels. Thus, 10% was

used as our benchmark value, but this still led to issues for the larger plantations, in which taking 10% of the pixels led to upwards of 3,000 pixels. It was decided that a hard cap of 20 pixels would be used for these bigger areas. The protocol for choosing the number of pixels to sample is visualized in Figure 4.6. The final export ended up taking around 3 days per country.

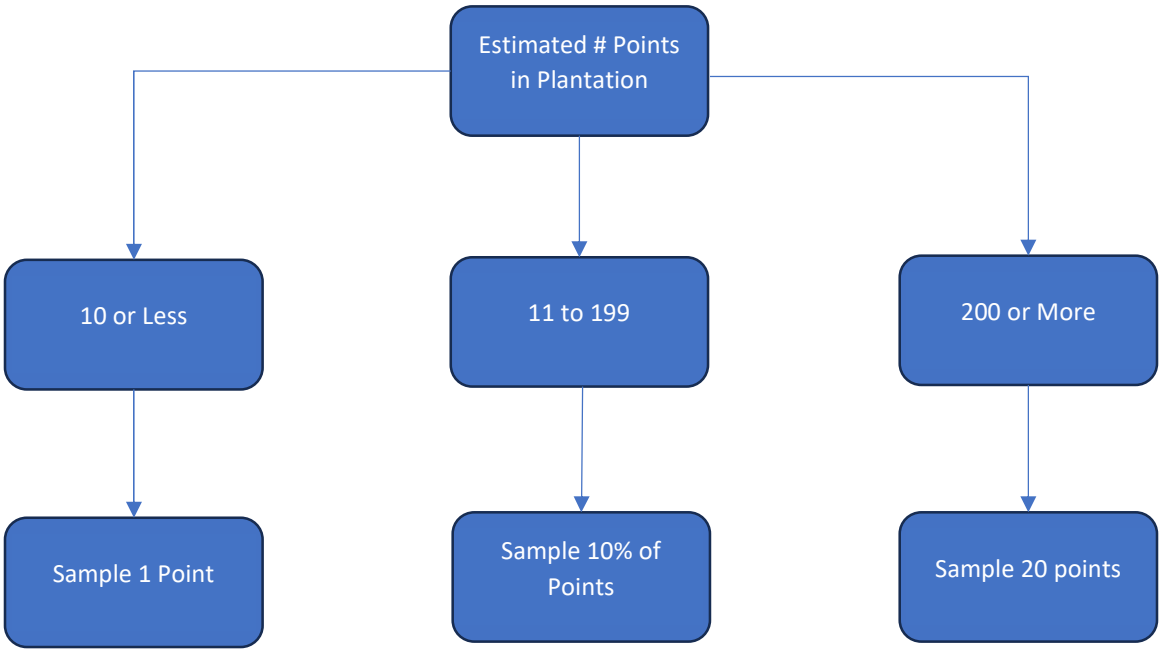


Figure 4.6 Sampling Process

4.2.3 Data Cleaning

The exported file containing the point measurements was very messy, having badly formatted columns along with suboptimal names. This subsection explores the data cleaning that was carried out to remedy these issues.

The first step involved using the lsat_format_data() function. The original file had one column for the date, one column for the latitude longitude pair, and had untidy names for many of the other columns. For example, all column names were completely capitalized, the spectral band columns were listed with their abbreviation rather than their full name, and the latitude longitude pair column was named ".geo". Once cleaned the dataset contained a singular column for each year, day of the year, latitude, longitude, identifier, and spectral bands. The column names were also changed to be a more intuitive description of what was contained within. An example row of the cleaned dataset can be seen in Table 4.2.

Table 4.2 Example Landsat Dataset Row

Sample.id	Year	DOY	Latitude	Longitude	Blue
S_1	1984	365	-1.3	-47.9	.087

Red	NIR	Satellite	Water	Clouds
.234	.567	LANDSAT_5	1	47

The lsat_clean_data() function was then used to filter out measurements that exhibited signs of inaccuracy. This includes measurements that were taken when there was rain or major cloud cover. This filtering reduced the number of measurements by around 70% for both Brazil and Peru.

4.2.4 Calibrating EVI Measurements

The next step of the process involved calculating the EVI values from the spectral band measurements, followed by calibrating the EVI values from the various satellites, Landsat 5, 7, and 8. The calculation of EVI was straightforward, using the formula as listed in subsection 2.1.1.

Calibrating the EVI measurements is necessary because there are systematic differences between the instruments on the Landsat satellites used to measure the spectral indices. Landsat 5 was deployed from 1984 to 2013 and used Thematic Mapper and Multispectral Scanner instruments to measure spectral indices, Landsat 7 from 1999 to the present using an Enhanced Thematic Mapper Plus instrument to measure, and Landsat 8 from 2013 to the present using Operational Land Imager and Thermal Infrared Sensor instruments. In theory these satellites should get the same measurements if measuring the same pixel at the same point in time, but practically that is not the case. There are natural differences due to the instruments used for measuring the spectral bands, one satellite may perform better in slightly cloudy conditions while one may perform better in wet areas. By calibrating the measurements, we were putting every measurement on the same scale (same Landsat satellite scale).

The calibration is done with the help of the lsat_calibrate_rf() function. There is overlap in satellite deployment between the Landsat 5 and 7 satellites, and Landsat 7 and 8 satellites. Because of this, Landsat 7 was used as the baseline satellite. The calibration involved fitting two separate random forest models. One model predicting Landsat 7 EVI from Landsat 5 EVI, day of the year and pixel coordinates. And one model predicting Landsat 7 EVI from Landsat 8 EVI, day of the year, and pixel coordinates. The data used

to fit the models was the overlapping dates between the satellites where there were two separate EVI values. For example, for the model predicting Landsat 7 EVI from Landsat 5 EVI, the data used was the dates where there was an EVI measurement for both Landsat 7 and Landsat 5. The training test split was 75/25. These models were then used to predict Landsat 7 EVI values for the days in which there were only Landsat 5 or Landsat 8 measurements. For days when there were both Landsat 7 and Landsat 5 or 8 EVI measurements, the Landsat 7 EVI measurement was used.

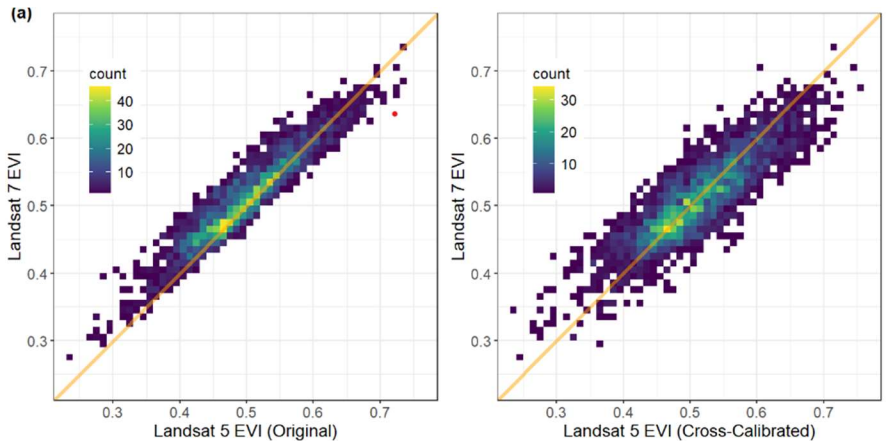


Figure 4.7 Landsat 5 EVI Calibration

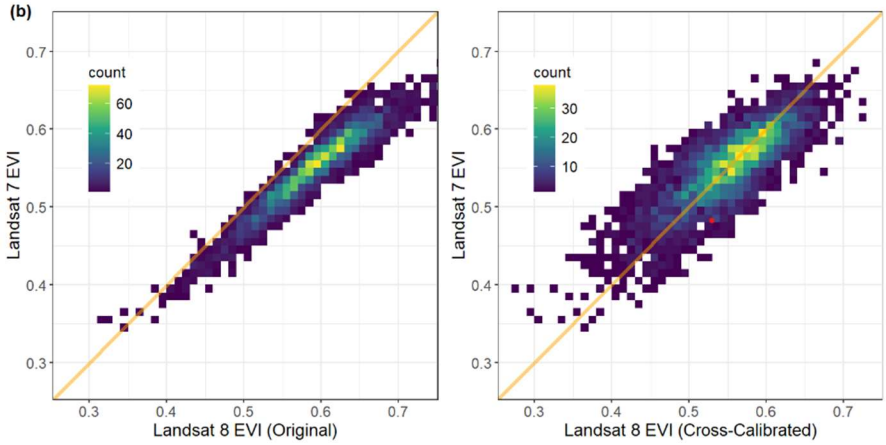


Figure 4.8 Landsat 8 EVI Calibration

Figure 4.7 depicts the Landsat 5 calibration process; the left graph plots original Landsat 5 EVI measurements vs the Landsat 7 EVI measurements for the same day. Ideally the points would fall in a straight line signaling the equality of the measurements between the two satellites. However, that was not the case, most points fell above the diagonal line meaning that Landsat 7 EVI values tended to be slightly larger than Landsat 5 values. The right graph displays the predicted (calibrated) Landsat 5 EVI values vs Landsat 7 EVI values (same values as the left graph). The points now fall more in line with the diagonal line, signaling that the model has done a good job adjusting the measurements to the Landsat 7 EVI scale.

4.2.5 Aggregating Points

At this point in the process there was data for EVI values at sample points along with information on the day and year that the measurement was taken. Each of the sample points belonged to a larger plantation, with the number of sample points from each plantation ranging from 1 to 20. Again, the goal of the analysis was to summarize the trends in EVI at each plantation to predict the deforestation year, thus sample points within the same plantation were aggregated to obtain one overall EVI measurement for each plantation at each point in time.

This aggregation was simply done by grouping sample points within each plantation by date and calculating the mean EVI value; this leaves one EVI measurement for each time point for each plantation.

4.2.6 Modeling EVI Over Time

Ideally there would be an EVI measurement for each day dating back to 1984 for each plantation. However, this was not the case due to the orbital nature of the satellite, taking 8-16 days for the satellite to return to the same point on Earth and weather conditions leading to inaccurate measurements. As stated, these measurements were filtered out of the data, leaving gaps in time. Thus, to get a clear picture of the trends in EVI, these gaps need to be filled in. This was done by fitting splines to the data.

For each plantation, a cubic spline was iteratively fit to aid in filling in the gaps of missing data. This results in a predicted EVI value for every day from 1984 to the present. The specificities of the spline fitting were as follows; the data was sorted by group and day of the year. Then, an initial cubic spline was fit to the observed EVI values for each plantation. These initial fits were used to predict fitted values for all dates. An iterative refitting process was then initiated if necessary, depending on the accuracy of the fits. If the percentage difference in predicted EVI versus observed EVI exceeded 30%, indicating a poor fit, the spline was refit. This process continued until a final curve was converged upon, leading to the final spline fit.

4.2.7 Identifying Deforestation Years

Each plantation had a predicted EVI value for every point in time. As mentioned, deforestation is signaled by the point in time with the minimum EVI value, the time when the plantation was the least green. To get the year when this occurred, the predicted EVI

values were grouped by year and the minimum EVI value was located for each plantation. This year was selected as the predicted year for when deforestation occurred.

4.2.8 Subsetting by Vegetation Type

As stated in subsections 4.2.1 and 4.2.2, the plantations in Peru are all currently palm oil plantations while the plantations within Brazil have varying primary vegetations, including palm oil, cocoa, and açaí. Because of this variation in Brazil, it was of interest to look at the deforestation years grouped by vegetation type. This was simply done by looking at a density plot of deforestation years grouped by vegetation type.

4.3 ALOS

This section aims to explain the steps to analyze deforestation trends in the plantations of interest, explaining the workflow for the analysis using measurements from the ALOS satellite. This involves using the HH and HV measurements from the satellite to make predictions on when deforestation occurred within the plantations. A big limitation with the ALOS satellite was that it was launched in 2015, meaning we only had access to measurements from 2015 to the present. Since the analysis is aimed at predicting deforestation year, and these predictions were solely based on the data from the satellite, we were only able to predict deforestation year as 2015 or more recently. This is an issue because the plantations in the study were deforested sometime between 1984 and 2024, the same range that was predicted from the Landsat analysis. To still make use of this satellite and its measurements, plantations used in the ALOS analysis were only those

that were predicted as being deforested from 2015 to 2024 by the Landsat analysis. This allowed the results from ALOS to be used as a confirmatory measure to the Landsat analysis, if the two analyses predicted the same year for deforestation, it is likely that year was truly when the plantation was deforested. This decreased the number of plantations used in the ALOS analysis to around 600 plantations.

4.3.1 Sample Point Selection and Export

The Google Earth Engine API was used through R to export HH and HV band measurements from the ALOS satellite for each of the plantations. There were 30 sample points taken from each plantation. Attempts were made to extract measurements for every pixel across each month spanning from 2015 to 2024, targeting a total of one measurement per month for each pixel. The export took around 5 hours for each Brazil and Peru.

4.3.2 Aggregating Points and HH to HV Ratio

The HH and HV band measurements were independently averaged within each plantation grouped by date. That is, the 30 sample points from each plantation were combined to get one HH and one HV value for each of the plantations for each point in time. An example row of the dataset after the averaging can be seen in Table 4.3.

Table 4.3 Example ALOS Dataset Row

Plantation	Date	HH	HV
1	2015-01-01	4678	2367

The ratio of HH to HV was calculated for each point in time that there was a measurement for HH and HV. The HH to HV ratio was used as the metric of interest for signaling deforestation, the reasoning for this explained further in following sections.

4.3.3 Modeling HH to HV Ratio Over Time

Ideally there would be an HH to HV ratio for every month dating back to 2015 for each plantation. However, this was not the case, certain time periods contained null values, meaning these values could not be read and were empty. This could be due to a variety of reasons. Imagery from the ALOS satellite is sometimes distorted due to extremely dense vegetation, which could be the case for certain regions in the Amazon. These null values could also come from general issues with the satellite itself such as maintenance. To obtain a clear picture of the trends in HH to HV ratio, these gaps needed to be filled in. This was done by fitting splines to the data.

For each plantation, a cubic spline was fitted to aid in filling in the gaps of missing data. This resulted in a predicted HH to HV value for every month from 2015 to the present.

4.3.4 Identifying Deforestation Years

Each plantation had a predicted HH to HV ratio for every point in time from 2015 to 2024. The point in time for each plantation in which the maximum HH to HV ratio was fit by the spline was selected as the deforestation year. This metric was used as the barometer for deforestation because HH measures the general roughness of the ground

surface and HV measures the vertical orientation of the vegetation, with taller vegetation typically having large values of HV. Thus, when deforestation occurs the HV should drop significantly while the HH increases due to the surface of deforested area becoming rougher. This increases the HH to HV ratio because it is likely that the numerator of the ratio increases while the denominator decreases.

Chapter 5

RESULTS

This section will report on the predicted deforestation years obtained from both the Landsat and ALOS satellites as well as the comparison between the results of the two. The plantations within Peru and Brazil will be reported on separately as they were treated as two separate groups during the analysis.

5.1 Peru

Before going into the subsequent results, it's important to note that while there were originally 1,281 plantations of interest in Peru, only 1,279 will be reported on. Two of the plantations were small enough that it was not possible to get accurate measurements, when exporting from the Google Earth Engine API no measurements were returned.

5.1.1 Landsat Deforestation Predictions

To reiterate how the predictions of deforestation year were made from the Landsat satellite, splines were fit to the time series of Enhanced Vegetation Index measurements. For each of the plantations, the year in which the minimum EVI value was fit by the spline was marked as the year of deforestation. This is because EVI is measuring the greenness of the vegetation within each specified plantation, so the lowest EVI

measurement is very likely when the plantation was deforested. This thought process is shown in Figure 5.1. The plot displays EVI information for a specific plantation, plantation 1,036. Each curve/spline represents a year and displays the fitted EVI values over time. There is a harsh drop in EVI around day 300 for the year 2008, this is the red curve that drops below an EVI of .5, this year is marked as the predicted deforestation year. This process is repeated for all the plantations to obtain predicted deforestation years. While the minimum EVI value is the main point of interest, this plot also gives an overall view of trends in EVI.

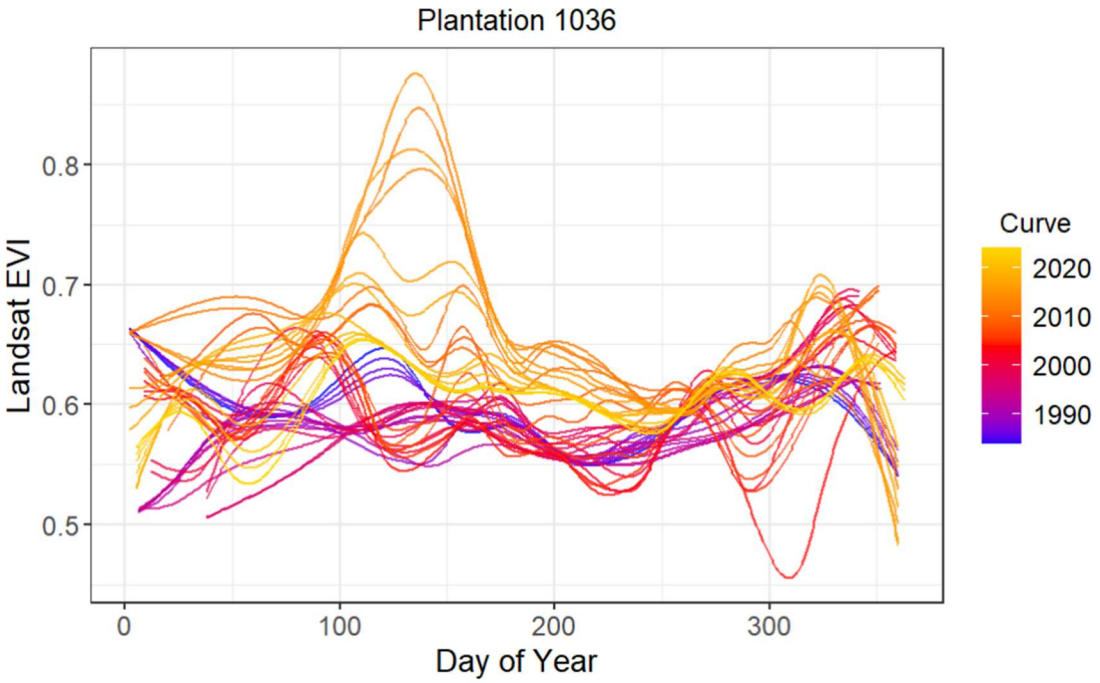
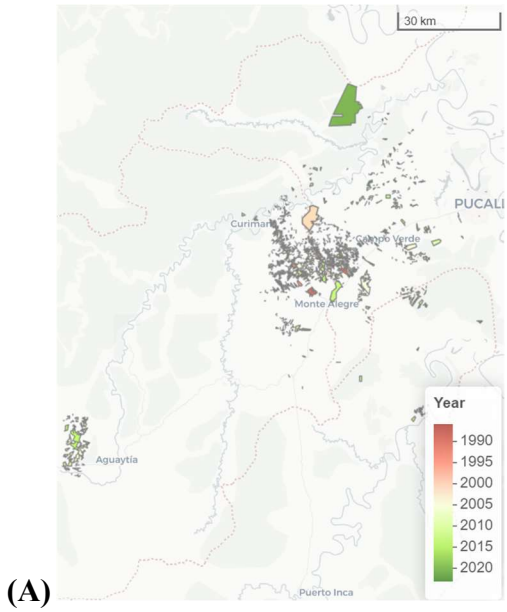


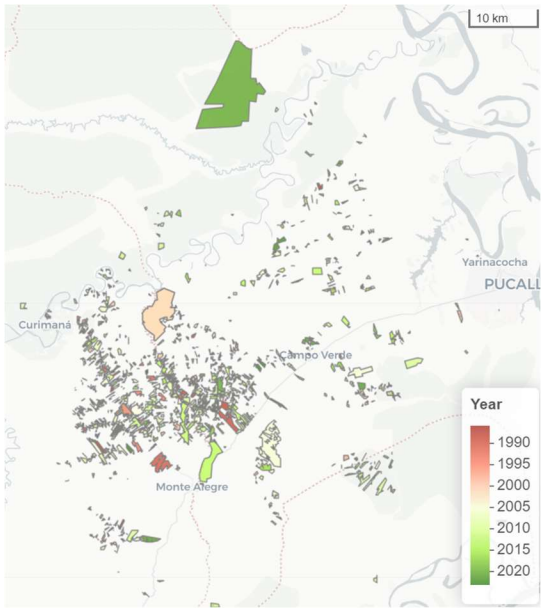
Figure 5.1 EVI Splines for Plantation 1036

The main contribution of the predictions is an interactive map displaying each plantation shaded by the year that it was deforested as seen in Figure 5.2. The shading is done based on a continuous spectrum, plantations predicted to be deforested closer to 1984 are shaded redder, while plantations predicted to be deforested closer to 2024 are shaded

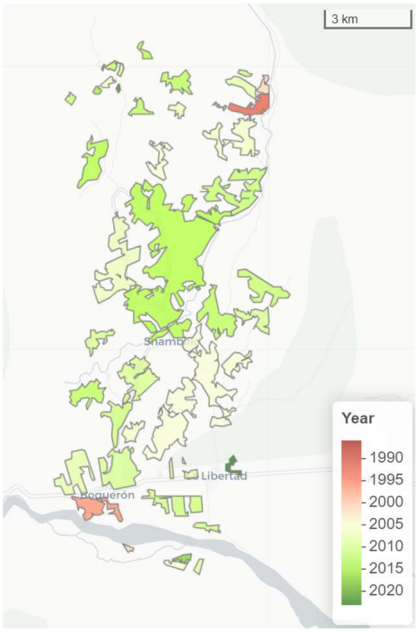
greener. There are three images displayed, one zoomed out image of all the plantations and 2 subsequent images zoomed in on different parts of the map.



(A)



(B)



(C)

Figure 5.2 Peru Landsat Prediction Map: Predicted deforestation year for plantations in Peru. Plot (A) includes all Peru plantations. Plot (B) includes plantations in the top center of plot A. Plot (C) includes plantations in the bottom left of plot A.

Table 5.1 Peru Predicted Deforestation Percentages

Deforestation Years	Percentage
1984-1993	5.2%
1994-2003	12.9%
2004-2013	62.3%
2014+	19.6%

Grouping the predicted deforestation years into bins of 10 years, a large proportion of the predicted deforestation happened in the mid-2000s, with over 60% of the predicted deforestation years being between 2004 and 2013. This is consistent with other sources which have tracked large amounts of deforestation in the Amazon related to the storm El Nino [11]. A density plot of the predicted years can be found in Figure 5.3. Reflecting the results from Table 5.2, a very large proportion of predicted deforestation years fall between 2002 and 2018.

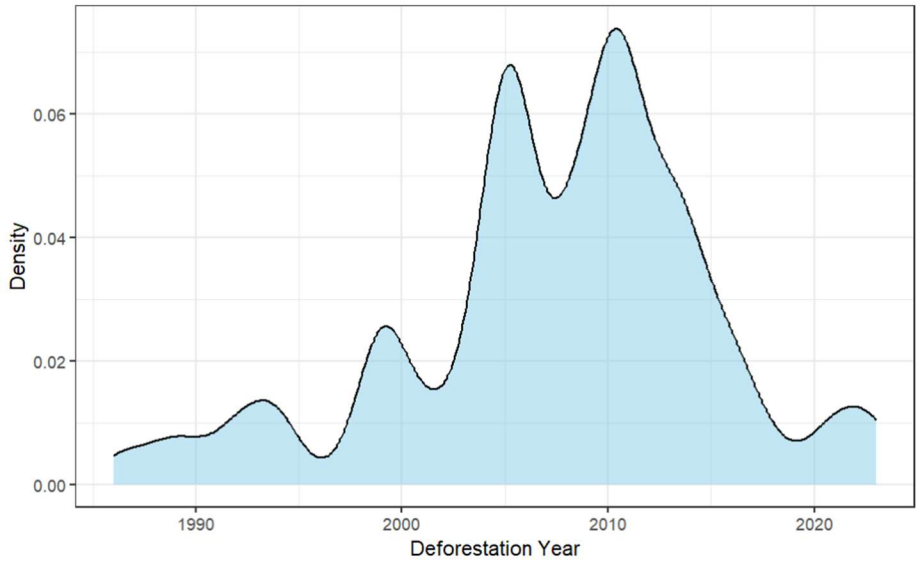


Figure 5.3 Peru Predicted Landsat Deforestation Year Density Plot

5.1.2 ALOS Deforestation Predictions

To refresh on the process for predicting deforestation year from the ALOS satellite, splines were fit to HH to HV ratio measurements within each plantation dating back to 2015. The year which contained the maximum predicted HH to HV ratio was selected as the year of deforestation due to the upward spike in ratio likely causing a drop in HV and/or an increase in HH, phenomena that have been linked to deforestation. An example spline fit for plantation 1159 is shown in Figure 5.4.

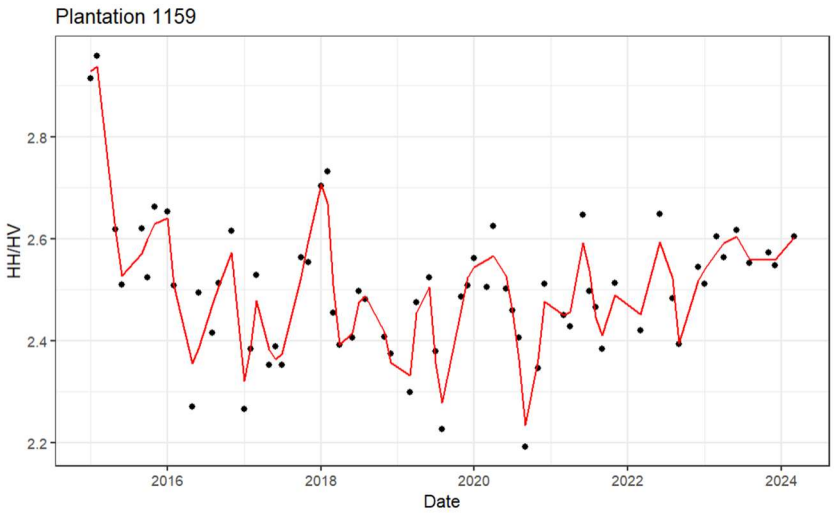


Figure 5.4 HH / HV Ratio Spline for Plantation 1159

The maximum predicted HH to HV ratio fit by the spline was in 2015, thus 2015 was predicted as the deforestation year for plantation 1159. Another contribution of the ALOS analysis is an interactive map displaying the predicted deforestation years for each plantation, as seen in Figure 5.5. Remember that ALOS deforestation predictions were made for every plantation of interest that was predicted to have a deforestation year of 2015 or later according to the Landsat analysis. This is why the range of shading is all green in Figure 5.5.

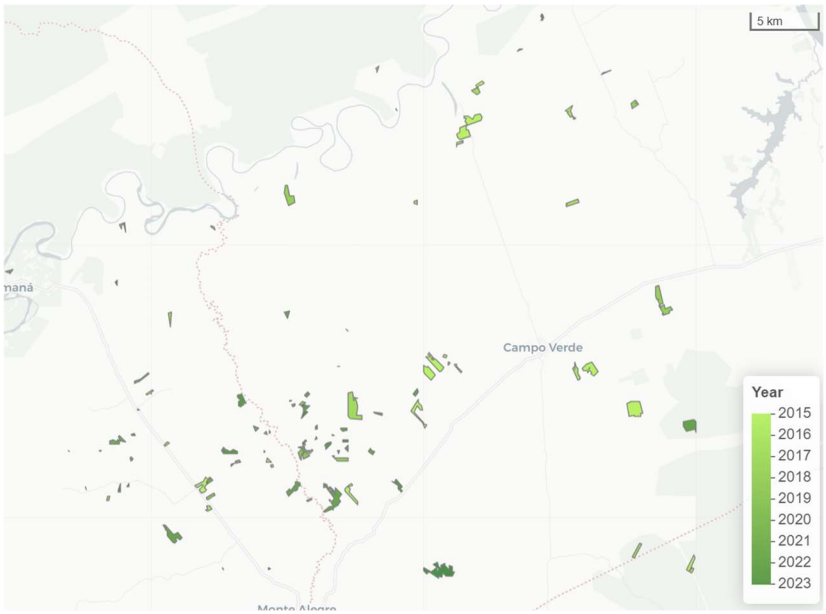


Figure 5.5 Peru ALOS Prediction Map

5.1.3 Landsat vs ALOS

This subsection is focused on comparing the predictions from the Landsat analysis to those of the ALOS analysis. Since ALOS was only able to predict deforestation years from 2015 on, plantations included for comparison are only those that were predicted to be deforested sometime 2015 and later by Landsat. The main form of comparison is looking at the difference in predicted deforestation year between the two analyses. These results are demonstrated in Figure 5.6. Around 13% of the plantations had the same predicted deforestation year for both the Landsat and ALOS analyses and around 25% within one year of each other. While there are a large proportion of plantations that had similar predicted deforestation years between the two analyses, there still were plantations that were very far off, with some plantations predicted deforestation years being more than 8 years different.

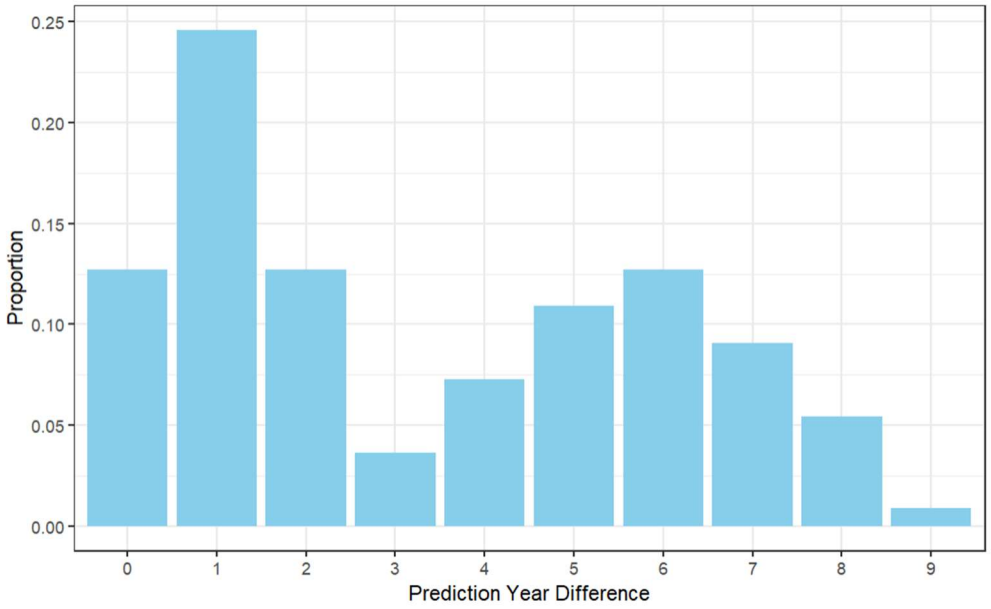


Figure 5.6 Comparing Predicted Deforestation Years Peru

5.2 Brazil

There were no issues of plantations being too small for the areas of interest within Brazil, so we have results for all 948 plantations. Since there was more crop diversity in the Brazil plantations than in Peru, there will be a section discussing the predictions subset by vegetation type.

5.2.1 Landsat Deforestation Predictions

The predictions for the Brazil plantations were done the same as they were for Peru, finding the year in the fitted splines where the EVI was at a minimum. Figure 5.7 displays four different views of the prediction map. The first, a zoomed-out image displaying all plantations, notice there are 3 chunks of plantations. The remaining three

images each are zoomed in on each cluster of plantations. Again, these plantations are shaded from red to green spectrum in relation to their predicted deforestation year.

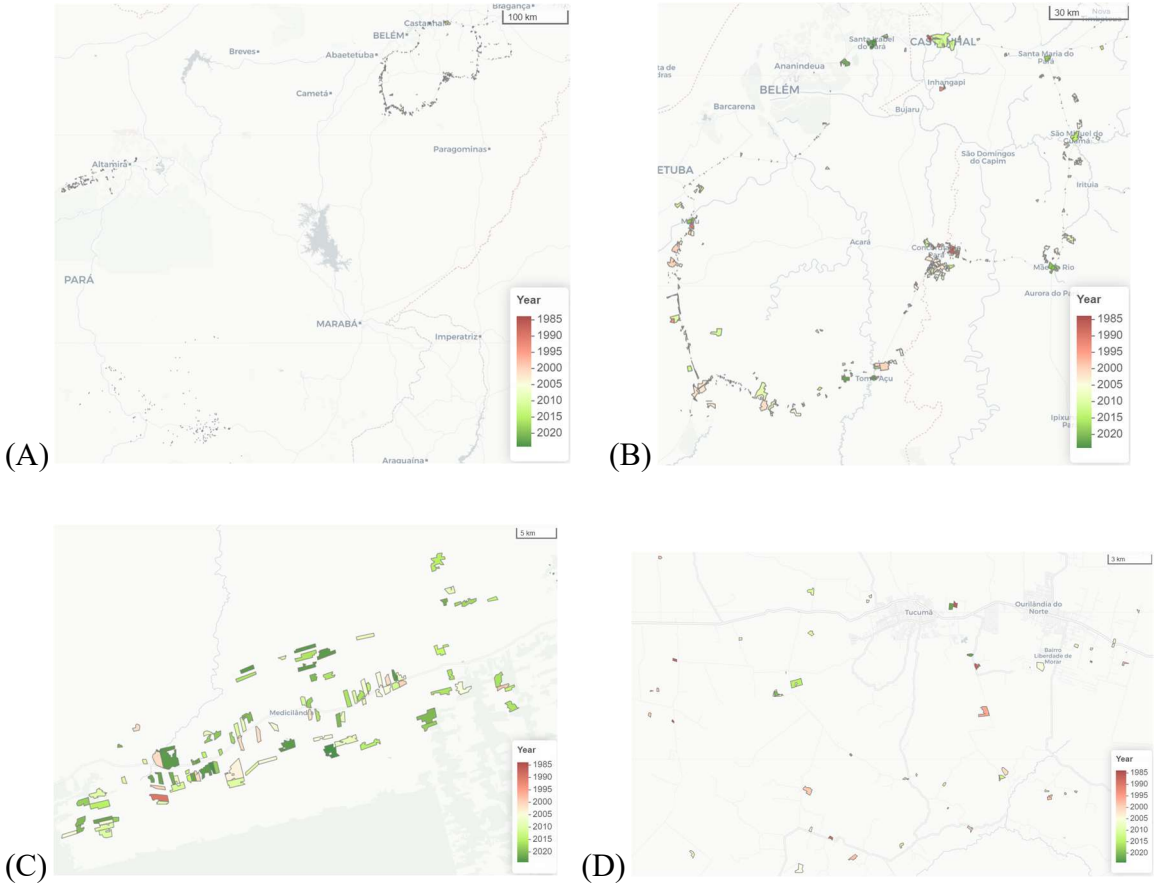


Figure 5.7 Brazil Landsat Prediction Map: Predicted deforestation year for plantations in Brazil. Plot (A) includes all Brazil plantations. Plot (B) includes plantations in the top right of plot A. Plot (C) includes plantations in the middle left of plot A. Plot (D) includes plantations in the bottom left of Plot A

Table 5.2 Brazil Predicted Deforestation Percentages

Deforestation Years	Percentage
1984-1993	5.9%
1994-2003	15.7%
2004-2013	41.1%
2014+	37.3%

A large proportion of the predicted deforestation years lay between 2004 and 2013 with an equally large proportion being between 2014 and later. This is similar to the distribution of predicted years in Peru in that not many plantations are predicted to be deforested before 2000. A density plot of the predicted years can be found in Figure 5.8. Mirroring the sentiment from Table 5.2, almost all the predicted deforestation years fall between 2000 and 2024 with 2005 to 2010 being the range in which the most deforestation years were predicted.

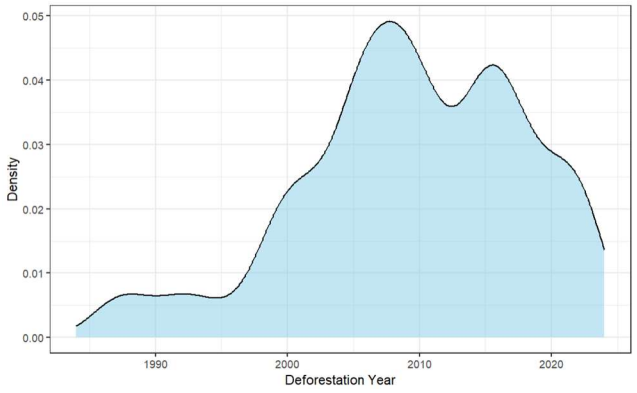


Figure 5.8 Brazil Predicted Landsat Deforestation Year Density Plot

5.2.2 Subsetting by Vegetation Type

The plantations in Brazil were primarily focused on one of three types of vegetation. These vegetation types include palm oil, cocoa, and açaí. The distribution of the Landsat predicted deforestation year grouped by vegetation type can be seen in Figure 5.9. The distributions are similar with the palm oil plantation predictions being the most different. These palm oil plantations have a much higher proportion of prediction years falling on

the lower end, before 1995, than the cocoa and açaí plantations and don't have many predictions after 2020. The distribution of prediction years for the cocoa and açaí plantations are almost identical in range and shape with the slight difference of the cocoa plantations' distribution being shifted a bit to the right.

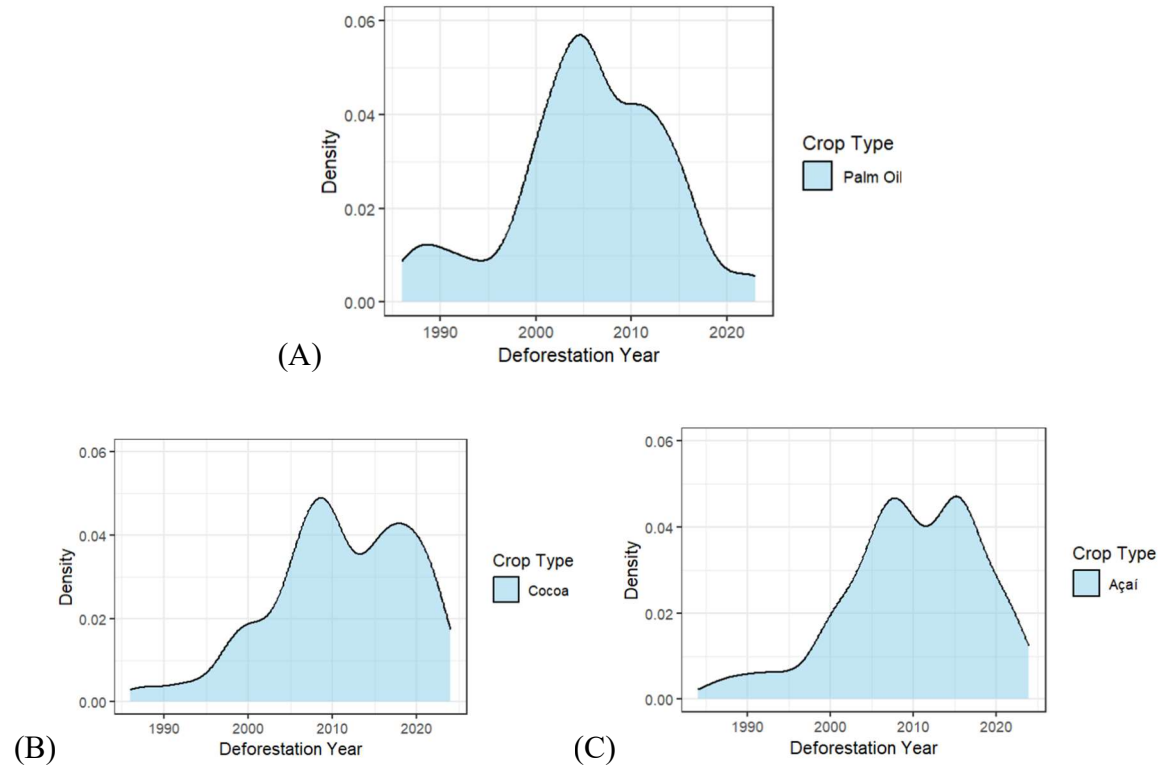


Figure 5.9 Predicted Landsat Deforestation Year by Vegetation: Plot (A) includes palm oil plantations. Plot (B) includes cocoa plantations. Plot (C) includes açaí plantations.

5.2.3 ALOS Deforestation Predictions

The predictions were done the same way for the ALOS analysis in Brazil as they were for Peru, finding the maximum year in which the spline is fit for each plantation. The results are displayed in Figure 5.10. Again, for the ALOS analysis, predictions were made only

for plantations which were predicted to have a deforestation year of 2015 or later for the Landsat analysis. This is reflected in Figure 5.10 as all the plantations are shaded green because the scale of shading is the same as it was for the Landsat analysis which went all the way back to 1984.

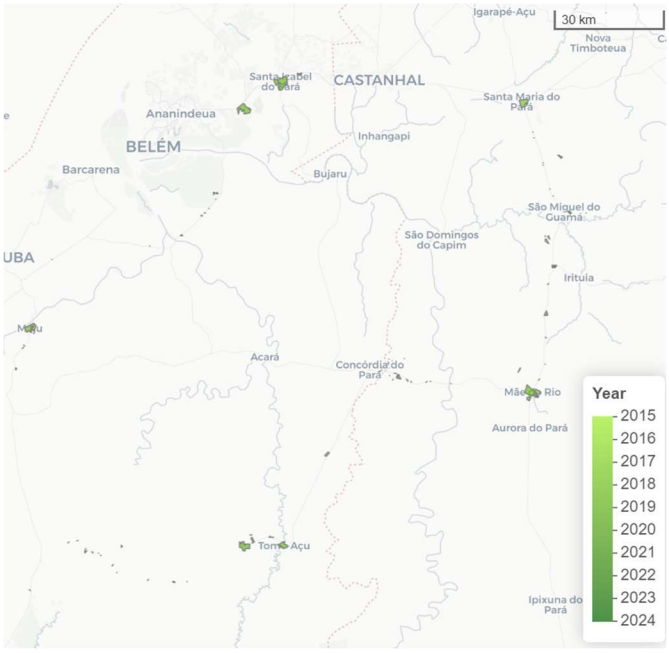


Figure 5.10 Brazil ALOS Prediction Map

5.2.4 Landsat vs ALOS

This subsection is focused on comparing the predictions from the Landsat analysis to those of the ALOS analysis for Brazil. Since ALOS was only able to predict deforestation years from 2015 on, plantations included for comparison are only those that were predicted to be deforested sometime 2015 and later by Landsat. The main form of comparison is looking at the difference in predicted deforestation year between the two analyses. These results are demonstrated in Figure 5.11. The two analyses seem to be

agreeing at some level, with around 50% of the plantations having predicted deforestation years within 2 years of each other between the two analyses. However, similarly to the Peru results, there are still many plantations that have predictions that disagree by 5+ years.

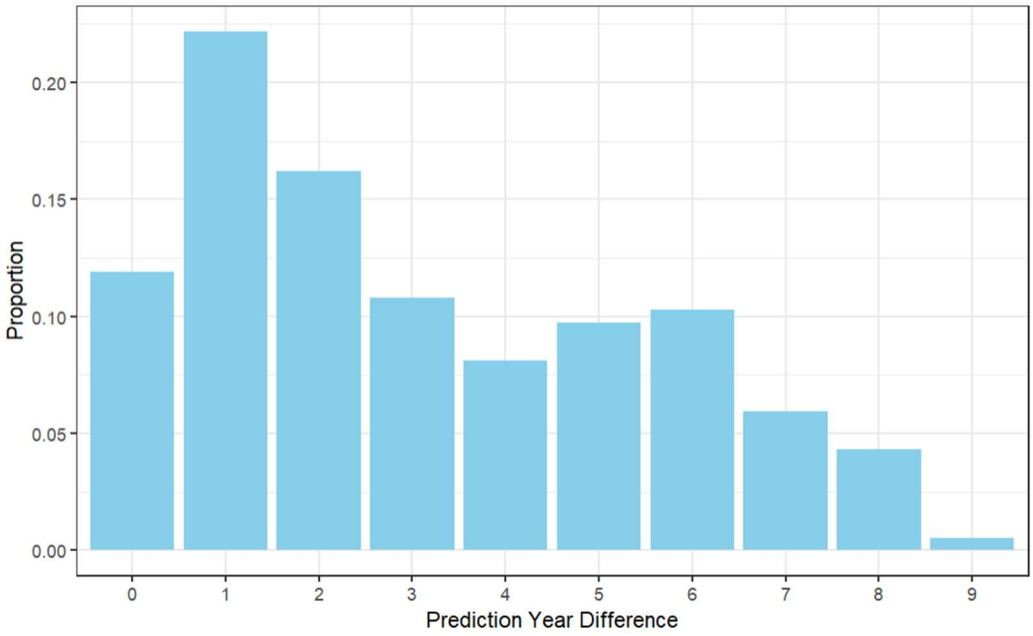


Figure 5.11 Comparing Predicted Deforestation Years Brazil

Chapter 6

CONCLUSION

This thesis explores two different methods for predicting deforestation year of a group of plantations in the Amazon Rainforest with the main goal of obtaining accurate predictions that can be used to trace land cover history of the areas of interest dating back to 1984.

For the first method, using the Landsat satellite, Enhanced Vegetation Index measurements were extracted for pixels within the plantations of interest with the help of the Google Earth Engine API. These EVI measurements were then cross calibrated using a random forest model, followed by splines being fit for each plantation to fill in the gaps of missing data due to extraneous factors such as weather or even general issues with the Landsat satellite. The splines served as a vessel to the history of the plantations, with a predicted EVI value being made for every point in time for each plantation. The predicted deforestation year was then obtained by locating the year in which the minimum EVI was fit by the spline, signaling the point in time when the greenness of the plantation was at a minimum. The predicted deforestation years were visualized in a map containing an outline of each plantation, shaded a color dependent on the year in which deforestation was predicted. A majority of the predicted deforestation years across both countries, Peru and Brazil, were in the range of 2002 to 2018. In Brazil the results were grouped by main vegetation type with the predicted deforestation years being very similar across cocoa and

açaí plantations, with palm oil plantations tending to have earlier predicted deforestation years.

The second method, using the ALOS satellite, gave an alternative set of predicted deforestation years. The general workflow of the ALOS satellite was almost entirely identical to that of the Landsat analysis. The main differences being the use of a different metric to predict deforestation, a lack of a calibration step, and the scope of years we could predict deforestation. The ALOS satellite used in this analysis was launched at the end of 2014, so data was only available from this time on. The HH and HV spectral band measurements from the satellite were combined into a HH to HV ratio and used as the metric of interest to signal deforestation. Splines were fit to the HH to HV ratio measurements for each plantation, with the year of the maximum spline fit value of the ratio being marked as the predicted year of deforestation. The results were displayed in a map containing the outline of each plantation shaded based on the predicted year of deforestation.

The final step was comparing the results of the two analyses. As stated within the paper, only plantations that were predicted to have a deforestation year 2015 or later were included in this comparison. There was some agreement between the two methods. In Peru and Brazil around 50% of the plantations had a difference in predicted deforestation year between the two methods laying between 0 and 2. For Brazil, 11.2% of the plantations had the same predicted deforestation year for both the Landsat and ALOS methods, for Peru this number was 12.7%. By far the most common difference in prediction years was 1 year; with Brazil having 22.1% falling in this category and Peru

24.7%. There was also disagreement with the methods as around half of the plantations had differences in predictions falling between 3 to 9 years apart.

6.1 Discussion

The next hurdle to jump through is which predictions to trust. For the plantations where there are predictions from both methods, there is a variety of routes to take. The first being, only accepting the deforestation year as correct if there is complete agreement between the two methods for the specific plantation. This would mean there is a difference of prediction years of 0 between the two methods. An alternative route could be creating an uncertainty interval over when the deforestation occurred. For example, for a plantation which had a predicted deforestation year of 2017 from the Landsat analysis and 2019 from the ALOS, it could be reported that the deforestation happened sometime between 2017 and 2019. There could even be a margin of error accounted for in the same way that a confidence interval works.

For the plantations with predicted deforestation years before 2015, we only have predictions from the Landsat analysis. The options are either to take the Landsat prediction as accurate or create a confidence level in the prediction based on other characteristics of the plantation. This could be considering things such as the difference in EVI between the point located as having the absolute minimum and other local minimum EVI values.

6.2 Reflection

While this research provides insight into the land cover dynamics within plantations in the Amazon, it does have its limitations. One point of interest is that each plantation is aimed at representing one farm entity, but based on the size of some of the larger plantations of interest it is likely that the singular plantation is actually a conglomeration of plantations. This would mean that the largest plantations may need to be split into smaller parts before running the analysis on them. Another possible limitation is that due to the computational expense of exporting many pixels from the plantations using the Google Earth Engine API, a sample of pixels was taken from each plantation. While a sampling procedure was put in place that sampled a specified number of points based on the area of the plantations to make sure to capture a representative sample of the entire plantation, there was an upper limit. The max number of points we were able to take from each plantation was 20, while this suffices for most of the plantations, a select few were very large and 20 points accounts for less than 1% of the possible pixels that could be sampled. This could lead to an unrepresentative sample and invalid results for these very large plantations.

6.3 Future Work

Ideas for future work and improvement all stem from the limitations listed above, most of which involve issues with the larger plantations. First steps could involve finding a solution to the problem of the very large plantations possibly being made of multiple plantations that were deforested at different times. Another branch that can be explored further is figuring out a way to make the export of pixels work more efficiently. This

would allow more pixels, maybe even all pixels, to be sampled from each plantation allowing for a clearer picture of what is occurring within the plantations rather than having to rely on a sample.

Another idea like that listed in the previous paragraph, is identifying which time of year deforestation is carried out based on reading other literature. If deforestation is something that only occurs in a specific season within the Amazon, for example, the summer. Then only measurements from the summer would need to be extracted, thus limiting the number of points that would need to be extracted.

While the steps listed above will likely increase the agreement between the two methods of predicting deforestation, there can be more steps taken to refine the process of using the measurements from the satellites to predict the deforestation years. This could involve using a different metric than the minimum EVI for Landsat or maximum HH to HV ratio to signal deforestation.

Another beneficial step would be to find exactly when some number of these plantations of interest were deforested manually, by looking at satellite imagery directly. While this process would take a long time to do for all plantations, getting an exact year of deforestation for even 10 of these areas of interest could be used as validation data. Comparing our predictions to the accurate manually tracked deforestation year can help to signal whether our predictions are accurate.

The entirety of the code used for this research is contained within the following GitHub repository: Ryan-DeStefano/Amazon-Remote-Sensing (github.com)
The purpose of the repository is to allow for an easy extension of the research discussed in this thesis through possible improvements as mentioned above. Reading the README

file in the repository along with this thesis should give anyone a solid base to understand the research and expand upon it.

BIBLIOGRAPHY

[1] *Satellites | Landsat Science*. (2021, November 30). Landsat.gsfc.nasa.gov. https://landsat.gsfc.nasa.gov/satellites/

[2] *Spectral bands and the EM spectrum | Spectral analysis of imagery*. (2022, August 25). Bas.ac.uk. https://guides.geospatial.bas.ac.uk/spectral-analysis-of-imagery/spectral-bands-and-the-em-spectrum

[3] *Landsat Enhanced Vegetation Index | U.S. Geological Survey*. (n.d.). Www.usgs.gov. Retrieved May 12, 2024, from https://www.google.com/url?q=https://www.usgs.gov/landsat-missions/landsat-enhanced-vegetation-index&sa=D&source=docs&ust=1715550489247684&usg=AOvVaw0zH5lYa1DAKiEcnYgWLECt

[4] *ALOS-2 (Advanced Land Observing Satellite-2) / Daichi-2 - eoPortal*. (n.d.). Www.eoportal.org. Retrieved May 12, 2024, from https://www.eoportal.org/satellite-missions/alos-2#eop-quick-facts-section

[5] Hashemvand Khiabani, P., & Takeuchi, W. (2020). Assessment of palm oil yield and biophysical suitability in Indonesia and Malaysia. *International Journal of Remote Sensing*, *41*(22), 8520–8546. https://doi.org/10.1080/01431161.2020.1782503

[6] Vera-Vélez, R., Grijalva, J., & Cota-Sánchez, J. H. (2019). Cocoa agroforestry and tree diversity in relation to past land use in the Northern Ecuadorian Amazon. *New Forests*, *50*(6), 891–910. https://doi.org/10.1007/s11056-019-09707-y

[7] Pereira, L., Corina, S.J.S. Sant'Anna, & Mariane Souza Reis. (2016). ALOS/PALSAR Data Evaluation for Land Use and Land Cover Mapping in the Amazon Region. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. https://doi.org/10.1109/jstars.2016.2622481

[8] Walker, W. S., Stickler, C. M., Kellndorfer, J. M., Kirsch, K. M., & Nepstad, D. C. (2010). Large-Area Classification and Mapping of Forest and Land Cover in the Brazilian Amazon: A Comparative Analysis of ALOS/PALSAR and Landsat Data Sources. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *3*(4), 594–604. https://doi.org/10.1109/jstars.2010.2076398

[9] Berner, L. T., Assmann, J. J., Normand, S., & Goetz, S. J. (2023). "LandsatTS": an R package to facilitate retrieval, cleaning, cross-calibration, and phenological modeling of Landsat time series data. *Ecography*, *2023*(9). https://doi.org/10.1111/ecog.06768

[10] Reygadas Langarica, Y., Spera, S., Galati, V., Salisbury, D. S., Souza da Silva, S., & Novoa, S. (2021). Mapping forest disturbances across the Southwestern Amazon: tradeoffs between open-source, Landsat-based algorithms. *Environmental Research Communications*. https://doi.org/10.1088/2515-7620/ac2210

[11] Rojas, E., Zutta, B. R., Velazco, Y. K., Montoya-Zumaeta, J. G., & Salvà-Catarineu, M. (2021). Deforestation Risk in the Peruvian Amazon Basin. *Environmental Conservation*, *48*(4), 1–10. https://doi.org/10.1017/s0376892921000291

[12] Watanabe, M., Koyama, C. N., Hayashi, M., Nagatani, I., Tadono, T., & Shimada, M. (2021). Refined algorithm for forest early warning system with ALOS-2/PALSAR-2 ScanSAR data in tropical forest regions. *Remote Sensing of Environment*, *265*, 112643. https://doi.org/10.1016/j.rse.2021.112643

[13] Berner, L. (2024, May 8). *logan-berner/LandsatTS*. GitHub. https://github.com/logan-berner/LandsatTS

[14] DeStefano, R. (2024, May 2). *Ryan-DeStefano/Amazon-Remote-Sensing*. GitHub. https://github.com/Ryan-DeStefano/Amazon-Remote-Sensing

[15] Huete, A.R., et al. "The Use of Vegetation Indices in Forested Regions: Issues of Linearity and Saturation." *IGARSS'97. 1997 IEEE International Geoscience and Remote Sensing Symposium Proceedings. Remote Sensing - a Scientific Vision for Sustainable Development*, 2022, https://doi.org/10.1109/igarss.1997.609169

[16] "Landsat Collections in Earth Engine | Earth Engine Data Catalog." *Google Developers*, developers.google.com/earth-engine/datasets/catalog/landsat

[17] Aybar, Cesar. "Rgee: R Bindings for Calling the "Earth Engine" API." *GitHub*, 3 June 2024, github.com/r-spatial/rgee

[18] Rosen, Paul A, et al. "The NASA-ISRO SAR (NISAR) Mission Dual-Band Radar Instrument Preliminary Design." *International Geoscience and Remote Sensing Symposium*, 23 July 2017, https://doi.org/10.1109/igarss.2017.8127836

[19] *About SERVIR - NASA*. www.nasa.gov/servir/about-servir/

APPENDICES

Appendix A

TABLES

Table A.1 Predicted Deforestation Year Dataset Example Rows

Plantation ID	Minimum EVI	Year
1	.406	2011
2	.315	2012
3	.353	1998

Table A.2 Brazil Number of Plantations by Crop Type

Crop Type	# of Plantations
Cocoa	393
Açaí	244
Palm Oil	179

Table A.3 Peru Landsat vs ALOS Prediction Year Differences

Difference	Percentage
0	11.82%
1	22.73%
2	8.18%
3	10.91%
4	8.18%
5	10%
6	13.64%
7	7.27%
8	6.36%
9	0.91%

Table A.4 Brazil Landsat vs ALOS Prediction Year Differences

Difference	Percentage
0	11.89%
1	22.16%
2	16.22%
3	10.81%
4	8.11%
5	9.73%
6	10.27%
7	5.95%
8	4.32%
9	0.54%

Appendix B

CODE

Listed below are some of the shorter functions used throughout the analysis. Not all functions and code are listed. The entirety of the code for this research is present in the repository: Ryan-DeStefano/Amazon-Remote-Sensing (github.com), this is the same repository listed in the future work section of this paper.

B.1 Landsat Extract Plantation Function

```
extract_group <- function(sample_id, group_lengths) {
    """
    Assigns groups (plantation numbers) to sample points based on sample IDs and group lengths.

    Parameters:
    sample_id (vector of strings): The list of sample IDs in the format 'S_#'.
    group_lengths (vector of integers): The list of lengths of each group.

    Returns:
    vector of integers: The assigned group (plantation) for each sample point.
    """
   group_numbers <- as.numeric(sub("S_", "", sample_id))
    cumulative_lengths <- c(0, cumsum(group_lengths))
```

```
  higher_order_group <- rep(NA, length(sample_id))


  for (i in 1:length(cumulative_lengths)) {
    if (i == 1) {
      higher_order_group[group_numbers <= cumulative_lengths[i]] <- i
    } else {
      higher_order_group[group_numbers > cumulative_lengths[i - 1] &
group_numbers <= cumulative_lengths[i]] <- i
    }
  }
  return(higher_order_group)
}
```

B.2 ALOS Extract Plantation Function

```
extract_group <- function(sample_id, group_ids) {
  """
  Assigns groups (plantation numbers) to sample points based on sample IDs and group
lengths.

  Parameters:
  sample_id (vector of strings): The list of sample IDs in the format 'S_#'.
  group_ids (vector of integers): The list of group IDs used for the ALOS analysis

  Returns:
```

```
    vector of integers: The assigned group (plantation) for each sample point.
    """
  group_number <- as.numeric(sub("S_", "", sample_id))
  group_index <- ((group_number - 1) %/% 30) %% length(group_ids) + 1
  return(group_ids[group_index])
}
```

B.3 Calculate Sample Size Function

```
num_sample_points <- function(dataset) {
  """
    Calculates the number of sample points to take from each plantation based on the
        estimated number of pixels in the plantation

    Parameters:
    dataset (data.frame): A data frame with at least one column named 'estimated_points'
        containing numeric values.

    Returns:
    vector of integers: A vector containing the calculated number of sample points for each
        plantation.
    """
  num_points_vector <- c()

  for (i in 1:nrow(dataset)) {
    estimated <- dataset$estimated_points[i]
```

```
    if (estimated < 10) {
      num_points <- 1
    } else if (estimated > 200) {
      num_points <- 20
    } else {
      num_points <- max(1, as.integer(estimated * 0.1))
    }
    num_points_vector <- c(num_points_vector, num_points)
  }
  return(num_points_vector)
}
```

B.4 Formatting Exported Landsat Measurements Function

```
lsat_format_data <- function (dt) {
"""
Preprocesses satellite data.

Parameters:
dt (data.table): The input data table containing satellite data.

Returns:
data.table: Preprocessed satellite data table.
"""
  dt <- data.table::data.table(dt)
```

```
setnames(dt, "sample_id", "sample.id")

colnames(dt) <- tolower(colnames(dt))

colnames(dt) <- gsub("_", ".", colnames(dt))

data.table::setnames(dt, "spacecraft.id", "satellite")

dt[, `:=`(date.acquired, data.table::as.IDate(date.acquired))]

dates <- as.POSIXlt(dt$date.acquired, format = "%Y%m%d")

dt$year <- dates$year + 1900

dt$doy <- dates$yday

coords <- stringr::str_extract(string = dt$.geo, pattern = "(?<=\\[).*(?=\\])")

coords <- matrix(unlist(strsplit(coords, ",")), ncol = 2,

                 byrow = T)

dt$latitude <- as.numeric(coords[, 2])

dt$longitude <- as.numeric(coords[, 1])

setkey(dt, "satellite")

lsat57.dt <- dt[c("LANDSAT_5", "LANDSAT_7")]

lsat57.bands <- c("blue", "green", "red", "nir", "swir1",

                  "swir2")

colnames(lsat57.dt)[which(colnames(lsat57.dt) %in% paste("sr.b",

                                         c(1:5, 7), sep = ""))] <- lsat57.bands

lsat57.dt <- lsat57.dt[, `:=`(sr.b6, NULL)]

lsat8.dt <- dt["LANDSAT_8"]

lsat8.bands <- c("ublue", "blue", "green", "red", "nir",

                 "swir1", "swir2")
```

```
  colnames(lsat8.dt)[which(colnames(lsat8.dt) %in% paste("sr.b",
                                        c(1:7), sep = ""))] <- lsat8.bands
  dt <- rbind(lsat57.dt, lsat8.dt, fill = T)
  keep.bands <- c("blue", "green", "red", "nir", "swir1",
                  "swir2")
  dt[, `:=`("ublue", NULL)]
  scaled.bands.dt <- dt[, ..keep.bands] * 2.75e-05 - 0.2
  dt[, `:=`((keep.bands), scaled.bands.dt)]
  dt <- data.table::setnames(dt, "max.extent", "jrc.water")
  keep.cols <- c("sample.id", "latitude", "longitude", "jrc.water",
                 "satellite", "year", "doy", "collection.number", "landsat.scene.id",
                 "processing.level", "sun.elevation", "qa.pixel", "qa.radsat",
                 "cloud.cover", "geometric.rmse.model", "blue", "green",
                 "red", "nir", "swir1", "swir2")
  dt <- dt[, keep.cols, with = F]
  dt <- dt[order(sample.id, year, doy, satellite)]
  return(dt)
}
```

B.5 Filtering Observations Due to Weather Function

```
lsat_clean_data <- function (dt, cloud.max = 80, geom.max = 30, sza.max = 60,
          filter.cfmask.snow = T, filter.cfmask.water = T, filter.jrc.water = T) {
"""
Filters satellite data based on various quality control parameters.
```

Parameters:

dt (data.table): The input data table containing satellite data.

cloud.max (numeric, optional): Maximum allowed cloud cover percentage (default is 80).

geom.max (numeric, optional): Maximum allowed geometric RMSE model (default is 30).

sza.max (numeric, optional): Maximum allowed solar zenith angle (default is 60).

filter.cfmask.snow (logical, optional): Whether to filter out observations flagged as snow by CFMask (default is TRUE).

filter.cfmask.water (logical, optional): Whether to filter out observations flagged as water by CFMask (default is TRUE).

filter.jrc.water (logical, optional): Whether to filter out observations flagged as water by JRC (default is TRUE).

Returns:

data.table: Filtered satellite data table.

"""

```
  dt <- data.table::data.table(dt)

  n.orig <- nrow(dt)

  dt[, `:=`(clear, mapply(clear_value, qa.pixel))]

  dt <- dt[clear == 1]

  if (filter.cfmask.snow == T) {

    dt[, `:=`(snow, mapply(snow_flag, qa.pixel))]

    dt <- dt[snow == 0]

  }

  if (filter.cfmask.water == T) {
```

```
  dt[, `:=`(water, mapply(water_flag, qa.pixel))]
  dt <- dt[water == 0]
 }
 if (filter.jrc.water == T) {
  dt[, `:=`(jrc.water, as.numeric(jrc.water))]
  dt <- dt[jrc.water == 0]
 }
 dt <- dt[cloud.cover <= cloud.max]
 dt <- dt[geometric.rmse.model <= geom.max]
 dt <- dt[90 - sun.elevation <= sza.max]
 dt <- dt[qa.radsat == 0]
 dt <- dt[blue > 0.005][green > 0.005][red > 0.005][nir >
  0.005]
 dt <- dt[blue < 1][green < 1][red < 1][nir < 1]
 n.final <- nrow(dt)
 n.removed <- n.orig - n.final
 print.msg <- paste0("removed ", n.removed, " of ", n.orig,
  " observations (", round(n.removed/n.orig * 100, 2),
  "%)")
 print(print.msg)
 return(dt)
}
```