**CSE 545 : Big Data Analytics - Team Project Report**
**SDG 15 : Life on Land (Deforestation in South America)**
**Team : SAP**

**Saif Vazir** (112072061) **Anuj Verma** (112504481) **Priyanka Bedarkar** (112046765)

## Introduction

Forests cover 30.7 percent of the Earth's surface and, in addition to providing food security and shelter, they are key to combating climate change, protecting biodiversity and the homes of the indigenous population. By protecting forests, we will also be able to strengthen natural resource management and increase land productivity.[8]

Currently, 13 million hectares of forests are being lost every year. The persistent degradation of drylands has led to the desertification of 3.6 billion hectares. Even though up to 15 % of land is currently under protection, biodiversity is still at risk. Deforestation and desertification - caused by human activities and climate change - pose major challenges to sustainable development and have affected the lives and livelihoods of millions of people.

SDG 15 addresses the conservation and sustainable use of forests, other terrestrial ecosystems and biodiversity, including halting desertification and land degradation and combating illegal trade in endangered species.

In this project we attempt to analyse the trends in forest gain and loss in the Southern America over the years 2012 to 2018. We also intend to cluster and study the regions with similar deforestation rates so that we could suggest them with similar measures to curb the forest loss. We have used four data pipelines - Spark , Hadoop, Pytorch and MapReduce along with the concepts of Deep Neural Networks, Linear Regression and Data parallelism to perform large data analysis.

## Background

Hansen et. al. [3] examined global Landsat data at a 30-meter spatial resolution. The result of their study provides Global Forest Change (GFC) map that estimates forest cover in year 2000, as well as gain and loss events that happened from 2000 to 2018 (as of version 1.6 of the map). The map has resolution 1296001 x 493200 pixels and a pixel size 30 metres. Since pixels are squares, a single pixel represents 30m x 30m = 900$m^2$ area. It should be noted that the GFC map captures any sort of plantations as forests that happen to be taller than 5 metres.

## Data

We use the Global Forest Cover dataset [1] provided by Hansen et. al. [3] for our project. This global dataset is divided into 10 x 10 degree tiles, consisting of seven files per tile. Out of these seven files, we concern ourselves with three :

1. Tree cover for 2000 (**treecover2000**) : Defined as all vegetation taller than 5m in height as seen in year 2000. This is encoded as a percentage per output grid cell, in the range 0–100.
2. Global forest cover gain 2000–2012 (**gain**) : Defined as the inverse of loss, or a non-forest to forest change. This is encoded as either 1 (gain) or 0 (no gain).

3. Year of gross forest cover loss event (**lossyear**) : Forest loss during the period 2000–2018, defined as a stand-replacement disturbance, or a change from a forest to non-forest state. This is encoded as either 0 (no loss) or else a value in the range 1–17, representing loss detected primarily in the year 2001–2018, respectively.

We downloaded individual 10 x 10 degree granules TIFF image, each of which is a high quality image of size more than a GB(on decompression). We limit our analysis to the South American region. There are 126 TIFF image files which span this region of interest, thus giving us a dataset of approximately 126 GB. This satisfies the Large Observation Criteria (option-1) and warrants the use of Big Data frameworks.

## Methods

There are two major tasks we attempted to address in this project viz. (i) tracking deforestation rate from 2013 - 2018, (ii) clustering the regions with similar deforestation rates. We have used four data pipelines - Spark , Hadoop, Pytorch and MapReduce along with the concepts of Deep Neural Networks, Linear Regression and Data parallelism to perform large data analysis.

### *1. Tracking deforestation rate from 2013 - 2018 :*

We loaded all the 126 TIFF files (treecover2000, gain and loss for years 2012 to 2018) into HDFS. We use load these TIFF images into Spark RDD and apply MapReduce functions according to the tasks as depicted below.

#### *1.1. Forest Gain vs Year :*

We then use the forest gain TIFF files, which have values 0's for no forest gain and 1's for forest gain. We load these files in Spark RDD using binaryFiles(). Further, we use a Mapper function to map the “forest gain filename” as the key to a value which is basically the sum over the entire gain image. Thus, this method counts the total forest gain in a particular image (particular $10^o$ latitude span, $10^o$ longitude span and year). Pipelines used for this are : Spark and Hadoop (HDFS). Pipelines used for this task are : Spark and Hadoop (HDFS).

#### *1.2. Forest Loss vs. Year :*

Next, we use the forest loss files, which have value 0 (no loss) or else a value in the range 1–17, representing loss detected primarily in the year 2001–2018. Similar to the above, we load these files in Spark RDD using binaryFiles(). We use Mapper function to map the “forest loss filename” to a value which is basically the count of cells having year-number the same as the one in the filename. Thus, this method counts the total forest loss in a particular image (particular $10^o$ latitude span, $10^o$ longitude span and year). We plot the total forest loss vs. year as depicted in the Results Section (Figure 1). Pipelines used for this task are : Spark and Hadoop (HDFS).

### *2. Forest Change Maps :*

We then use the treecover, forest gain and forest loss TIFF files. Each TIFF image (~ 1 GB) spans a region of $10^o$ in latitude and $10^o$ in longitude and has 40000 x 40000 pixels. We used PIL library and Spark in python to divide the large TIFF file into smaller tiles of 400 x 400 pixels size (spanning across $0.1^o$ latitude and $0.1^o$ longitude). We plot some of these smaller tiles. In the Results Section (Figure 2), treecover is represented as green pixels, deforestation is represented by red pixels. Black pixels represent no forest cover in that area. Pipelines used for this task are : Spark and Hadoop (HDFS).

### *3. Clustering the regions with similar deforestation rates :*

As stated in the Introduction Section, we intend to cluster and study the regions with similar deforestation rates so that we could suggest them with similar measures to curb the forest loss. For clustering the regions (locations), we need to have features which represent these regions. There are two ways to obtain these representative features, as explained in the following.

#### *3.1. Feature Extraction Using AlexNet :*

We use the forest loss files, load them in Spark RDD using binaryFiles(). Similar to what was stated in Methods Section 2 above, we divide this 40000 x 40000 pixels file into 10000 smaller tiles of 400 x 400 pixels each using the Mapper function. We now transform these 400 x 400 dimension images to 224 x 224 x 3 so as to input it correctly to a pretrained AlexNet model. We extract output from the penultimate layer of the AlexNet model. This output serves as “features” for the corresponding 400 x 400 small tile. We then normalize these features and try to cluster similar features together using K-Means Algorithm. Pipelines used for this task are : Spark, Hadoop (HDFS) and PyTorch. Concepts used for this task are : Neural Networks, K-Means Clustering and Data Parallelism.

As was suggested by the Professor in the feedback, CNN cannot perform well as a feature extractor in our case as the maps are binary (black / white) and have no distinguishable features in a small tile (400 x 400). This method can be applicable at a larger scale where bigger images can have more distinguishable features in terms of shapes or boundaries. But our current configuration prohibits us to follow this direction as the large image (40000 x 40000) will not fit into main memory. Therefore, we will use hand-crafted features to perform clustering and evaluate its performance.

#### *3.2. Hand-Crafted Features and K-Means :*

We use the forest loss files, load them in Spark RDD using binaryFiles(). Again, we pass this RDD through a Mapper function. But unlike the previous tasks where we retrieve 400 x 400 tiles as the value for the key, in this task, we retrieve the aggregated (summed) loss. So from the Mapper function, we get 0.1° latitude span, 0.1° longitude span, year and the corresponding loss value. Note that the latitude, longitude and year values are coming from the forest loss filenames, and the loss value is merely computed as the count of cell-values that match the year in the 400 x 400 tile. The tuple [latitude, longitude, year, loss] serves as the feature for the particular location.

We then normalize these features and try to cluster similar features together. We use K-Means algorithm for different cluster sizes ranging from 1 to 15. We compute the sum of squared errors to evaluate how good the clusters are. We then plot the error vs. cluster size and employ Elbow method to get the optimal cluster size. In the plot shown in Results Section (Figure 3), by Elbow method, we can visually see that the number of intrinsic clusters is 3.

Hence, we performed clustering with 3 clusters and extracted random 50 data points in these clusters. We then plotted these data points as shown in Results Section (Figure 4). After clustering the hand-crafted features for loss / deforestation over the year 2018, we can observe that the clusters are well defined and isolated. The x-axis represents the longitude (30W-80W) and the y-axis represents the latitude (00N-60S). We can infer that forests lying between latitude 00N-30S and longitude 30W-60W have the same kind of deforestation. Similarly, forests

of region (30S-60S, 55W-80W) have similar deforestation and forests of region (00N-25S, 60W-80W) have similar deforestation. These results can help us target specific regions with specific activities to curb deforestation. It can be inferred that the hand-crafted features were indeed representative of the forest loss in the respective locations, and that the clusters are well-formed, and that the results are satisfactory.

### *4. Nearest Neighbors in a Cluster :*

Once we have 3 clusters formed by method described above (Methods Section 3.2), we pick a point in cluster-3, and find its nearest 5 neighbors in the cluster. We get these 5 data points with [latitude, longitude, year, loss] as the hand-crafted features. We use the latitude, longitude and year to reverse map it to its corresponding 400 x 400 tile. We plot these 400 x 400 tiles for the obtained 6 points in cluster-3, in an attempt to observe similar deforestation patterns in those regions. Refer Figure 5 in Results Section. Red pixels denote deforestation activities carried out in 2018 in those particular areas. Also Refer Results Section Figure 1b, which depicts the red cluster having the highest deforestation rates (regions in 00N-30S and 30W-60W).

### *5. Regression to Predict Loss Value :*

We went a step further to predict the deforestation / loss by using multivariate regression. We used the [loss] as the label (predictor variable), and [latitude, longitude, year] as the features (independent variables). We used the samples obtained from splitting 40000 x 40000 forest loss file for year 2017 using the approach discussed above in the Methods Section 2. We used 80% samples for training a Ridge Regression model and predicted the loss values for the remaining 20% samples. We got MAE : 1298.33, MSE : 7686148.22, RMSE : 2772.390344706261.
We then observed the weights (correlation factors) of every feature to draw insights regarding which features really affect the deforestation. We inferred that there was a positive correlation (1.20114281) between latitude and deforestation, and negative correlation (-5.42087066) between longitude and deforestation.

## Results

**Figure 1 : Forest Loss vs Year**
(Refer Methods Section 1.2)
Forest loss was the highest during 2016

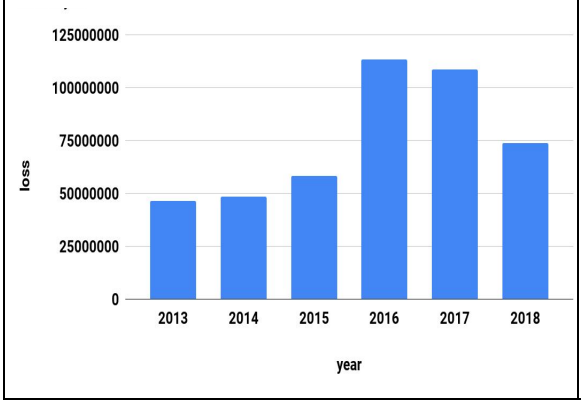



**Figure 1b : Cluster-wise trends of forest loss**
(Refer Figure 4 clusters and Methods Section 1)

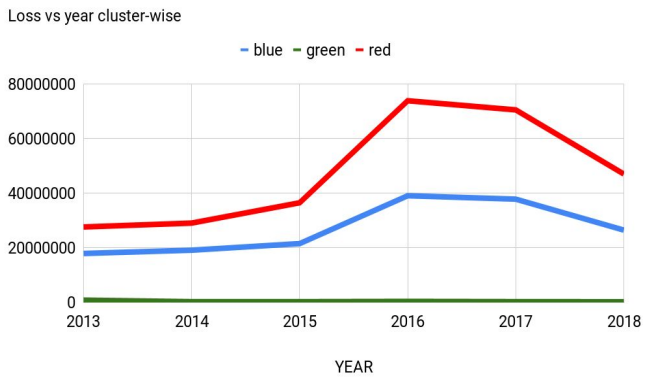


**Figure 2 : Forest Change Maps**
(Refer Methods Section 2)
Green pixels show tree cover, red pixels show
deforestation and black pixels represent no forest cover

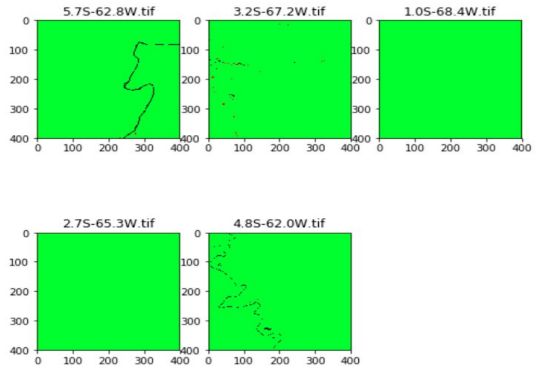



**Figure 1c : Forest Gain and Loss per 900m$^2$ in the 3 clusters obtained from K-means**
(Refer Methods Section 1)

| Cluster | Gain in 2012 | Loss in 2012 |
|---|---|---|
| Green | 5052474 | 785862 |
| Blue | 75552754 | 17872308 |
| Red | 202129591 | 27580910 |

**Figure 3 : K-Means Elbow Method**
(Refer Methods Section 3.2)
Cluster-size on x-axis vs Error value on y-axis. Observe that the optimal cluster size = 3.

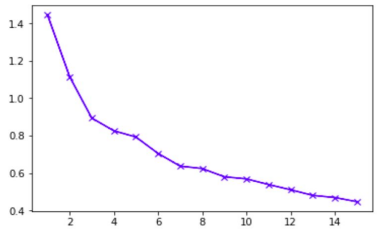


**Figure 4 : K-Means Clustering 3-clusters**
(Refer Methods Section 3.2)

| | | |
|---|---|---|
| 00N-30S | 30W-60W | red cluster |
| 30S-60S | 55W-80W | green cluster |
| 00N-25S | 60W-80W | blue cluster |

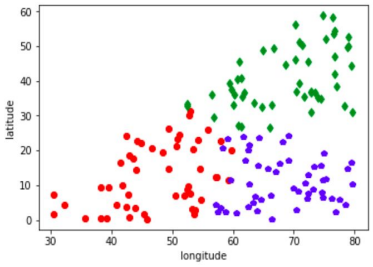
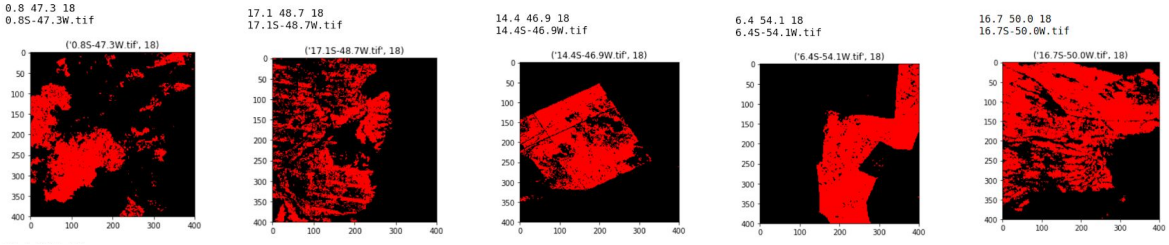



**Figure 5 : Deforestation Maps** (Refer Methods Section 4)
Red pixels denote deforestation activities carried out in 2018 in those particular areas



## Conclusion

1. Analysed Forest Gain trends across years 2012 to 2018.
2. Analysed Forest Loss trends across years 2012 to 2018.
3. Generated Forest change maps for smaller tiles.
4. Used CNN as Feature Extractor.
5. Used hand-crafted features and applied K-Means clustering, to locate clusters of regions with similar deforestation rates.
6. Used regression to find correlation between latitude, longitude and deforestation.
7. Used regression model to predict deforestation rates given latitude and longitude.

| Cluster | Location | Inference | Activities |
|---|---|---|---|
| Cluster-1 | North-Western Brazil | This cluster encompasses the Amazon basin present in the north-western region of Brazil. This has been the center of deforestation due to rich land available for agriculture. | Control slash-and-burn agriculture in this region to curb deforestation. Implement stricter regulations in regards to deforestation. |
| Cluster-2 | Bolivia, Peru, Chile | Rich tropical forests are cleared by the timber industry by logging. This has been the major reason in Bolivia and Peru. | Bolivia has already regulated it's forest resource usage, although the timber industry still makes a huge portion of Bolivia's exports. Hence, impose strict timber mining rules. |
| Cluster-3 | Northern Argentina | No major forests are present in this region. This cluster denotes region with less loss/gain activities. | We can promote reforestation activities in these areas as most of the land is under agricultural use. |

## References


1. http://earthenginepartners.appspot.com/science-2013-global-forest/download_v1.6.html
2. https://sustainabledevelopment.un.org/content/documents/196552018backgroundnotesSDG15.pdf
3. https://science.sciencemag.org/content/342/6160/850
4. https://www.lifegate.com/people/news/how-many-trees-in-the-world-how-many-we-cut-down
5. https://developers.google.com/earth-engine/tutorial_forest_02
6. https://unstats.un.org/sdgs/report/2016/goal-15/
7. https://sustainabledevelopment.un.org/content/documents/196552018backgroundnotesSDG15.pdf
8. https://www.un.org/sustainabledevelopment/biodiversity/