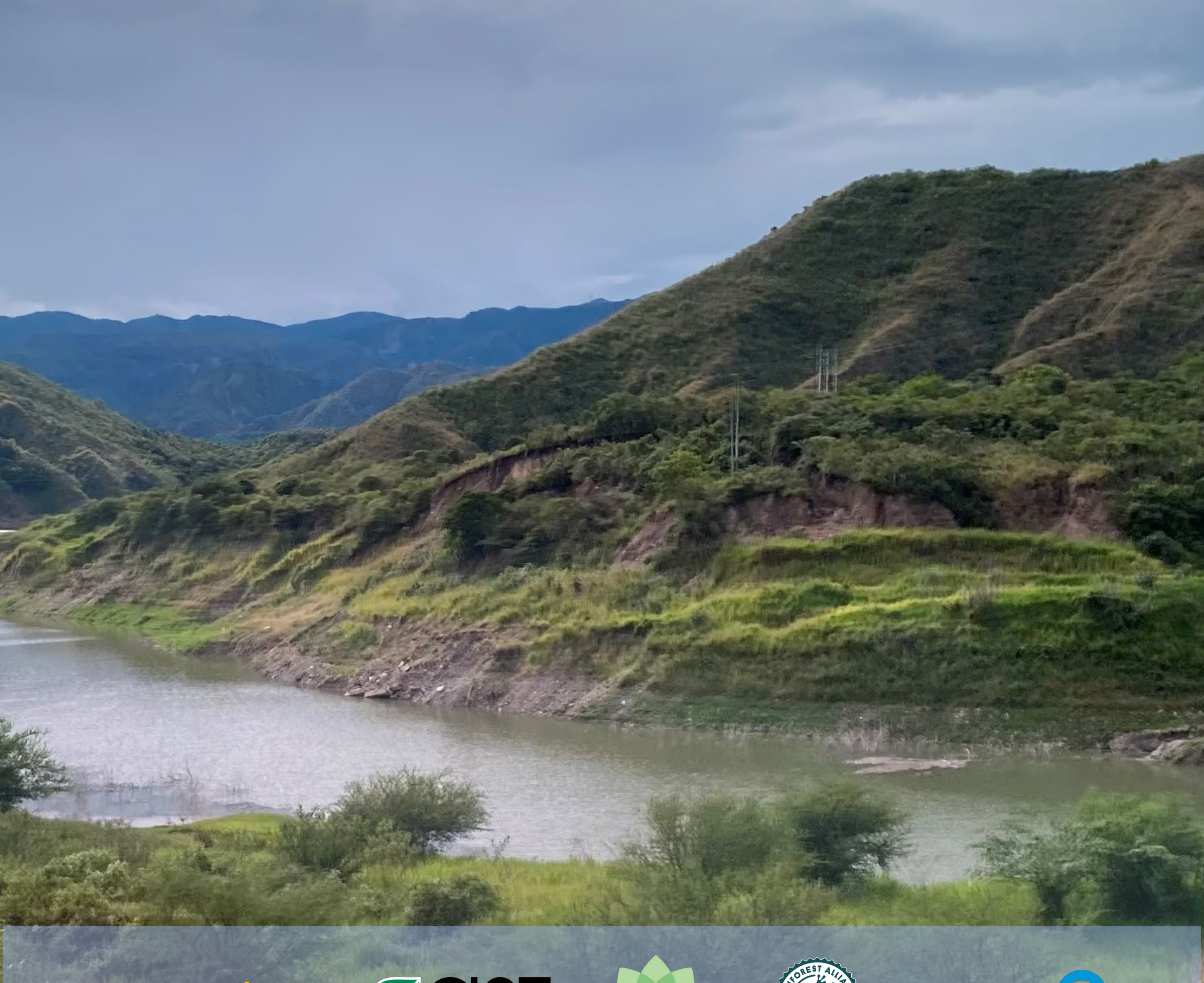
# METHODS & DATA SOURCES

Piloting improved assessment, monitoring and planning for deforestation free coffee landscapes in Colombia & Indonesia



# Table of Contents

## Section 1: COFFEE LAND USE MAP

- Colombia coffee land use map: Methodology
- Indonesia coffee land use map: Methodology

# COLOMBIA COFFEE LAND USE MAP: METHODOLOGY


***Authors***

Jorge Andres Perez Escobar. Research associate. Alliance of Bioversity and CIAT.
Fabio Alexander Castro. Research associate. Alliance of Bioversity and CIAT.
Paula Andres Paz. Research associate. Alliance of Bioversity and CIAT.
Louis Reymondin. Scientist, Terra-i Project coordinator. Alliance of Bioversity and CIAT.
Christian Bunn. Scientist, Climate Action. Alliance of Bioversity and CIAT.

***Image classification team***

Jorge Andres Perez Escobar. Research associate. Alliance of Bioversity and CIAT.
Fabio Alexander Castro. Research associate. Alliance of Bioversity and CIAT.



***Recommended citation***

Perez, J.; Castro, F.; Paz, P.; Reymondin, L; Bunn, C. (2021). Mapping coffee land use in Colombia for the year 2018. Alliance Bioversity & CIAT. Cali, Colombia.


## 1 Overview

This document provides a brief summary of the methods used to develop a map of coffee and non-coffee land uses for the year 2018 in Colombia. The method consists of four general steps: initial training data collection, satellite data pre-processing, algorithm training, data classification and results post processing, and fourth, results validation. The initial step consists of the compilation and cleaning of reference data of the land use classes coffee and other land uses in the form of polygons in Google Earth Pro (GEP), and satellite data collection from Sentinel-2 Level 2A (surface reflection), Sentinel-1 GRD, digital elevation model SRTM on Google Earth Engine (GEE). The satellite data is pre-processed to generate summary images for the year 2018. These data were used to train the RandomForest classifier to map coffee and non-coffee land uses. Spurious classifications were removed during post-processing. The final result was evaluated following Oloffson (2014) and a reference data set created from high resolution images in GEP. Validation demonstrated an 80.6% overall accuracy and a user accuracy of 72 and 85% for coffee and non-coffee land use respectively.

## 2 Data and methods

The methods used to create a map of coffee vs other land uses for Colombia consists of four main steps as shown in Figure 2. The first step is the collection of input data, the second the pre-processing of satellite images, the third the land use classification and, last, validation of the model with evaluation data.

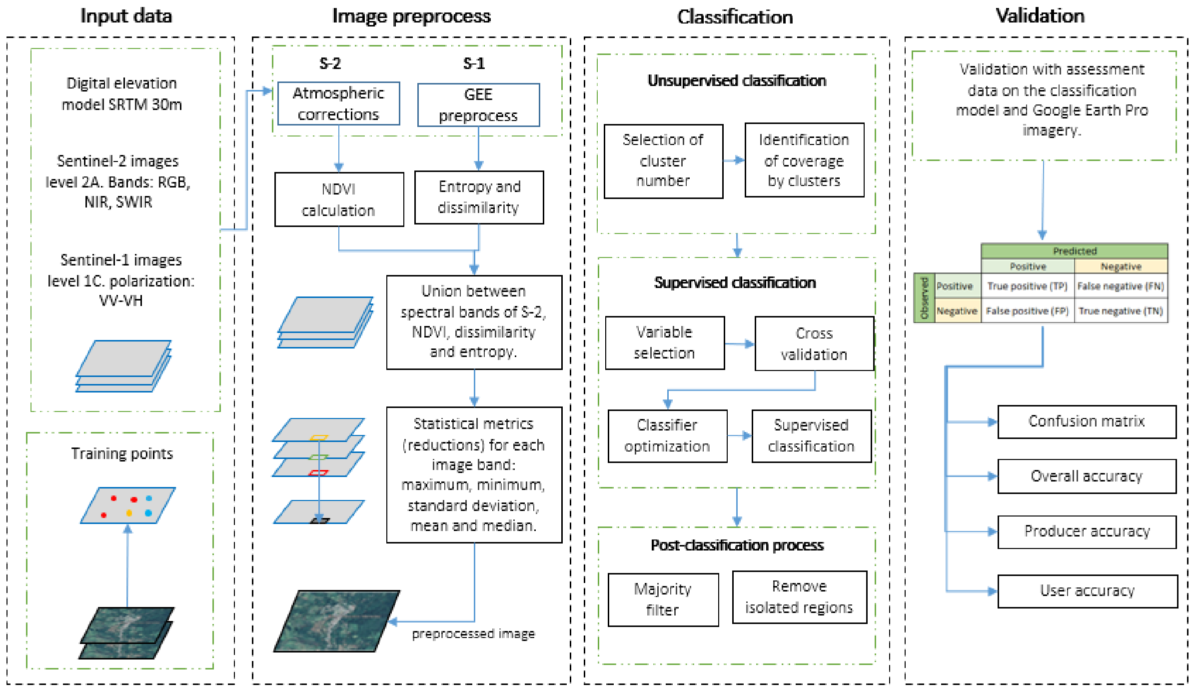



*Figure 1 Overview of methodology*

# 3 Input Data

Three types of information were used for the land use classification into coffee and no-coffee in Colombia. These were the creation of the zones of interest, satellite data, and training data. The first data set described the zones which were subject to classification into coffee and no-coffee. The second consists of the collection S2-2A and S1 with spatial resolution of 10m, and additionally the digital elevation model STRM of 30m, obtained through the data repository of Google Earth Engine (GEE). The third data set was created by collecting presence samples of coffee and no-coffee in the platform Google Earth Pro (GEP).

### *3.1.1 Region of interest*

The region of interest are areas which we can feasibly expect to contain coffee area. This data was constructed by using three geographical input data: a map of the climatic suitability for coffee in Colombia, the map of forest/no forest from the Instituto de Hidrografía, Meteorología y Estudios Ambientales – IDEAM for the year 2018, and lastly municipalities with relevant coffee area.

With this information, we delineated the region of interest by exclusion of a) pixels not climatically suitable for coffee, b) pixels with forest cover according to IDEAM, and c) municipalities which cumulatively include 5% of all harvested area of coffee in Colombia. The result can be seen in Figure 3.

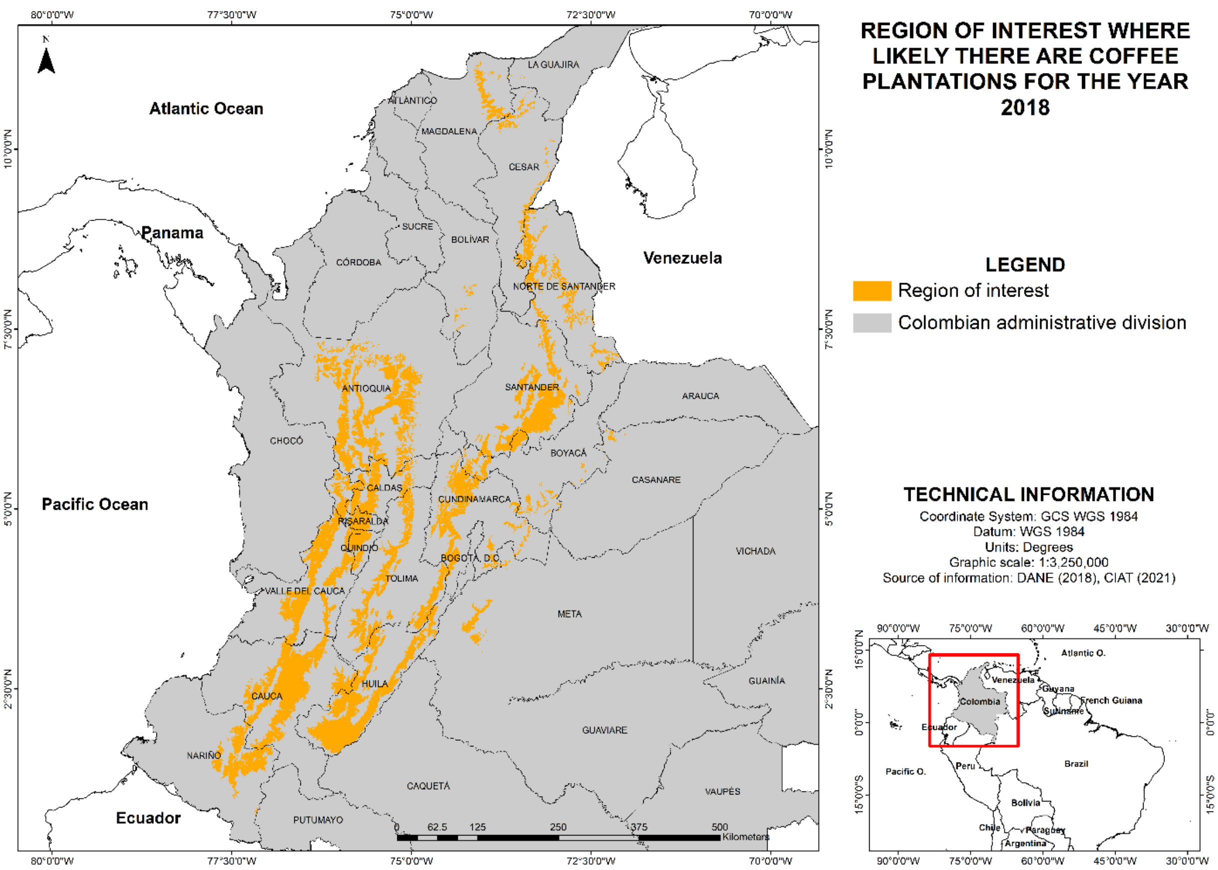



*Figure 2 Region of interest for classification; orange pixels can feasibly be expected to contain coffee area in 2018.*

In the following, we further describe the input data used for the zone of interest definition (Figure 5).

### *3.1.1.1 Map of suitable area for coffee*

We generated a suitability map for coffee in Colombia by classifying climate data with a Random Forest classifier. The classifier was trained with the raw presence data set as described below, and WorldClim climate data at 1km resolution. WorldClim provides monthly climate data for 1980-2010 ("current") from which we generated 19 bioclimatic variables. The variables describe seasonality and extremes of temperature and precipitation throughout the year. The RF classifier was trained on this data and extrapolated on the climate maps. The result was a map which shows the similarity of a location with the climate at current coffee locations. We assumed that locations with a suitability value lower than the 1$^{st}$ percentile at presence locations are climatically unfeasible for coffee production.

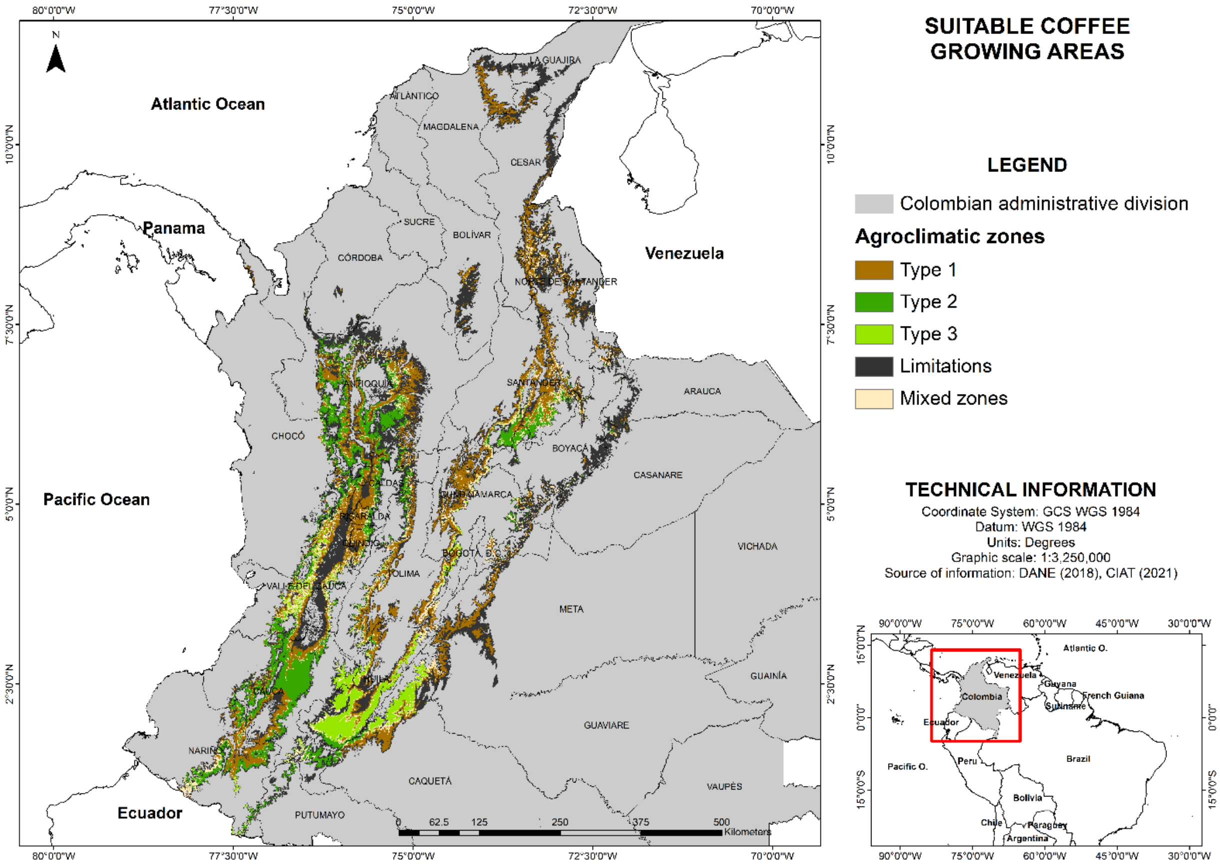



*Figure 3 Suitability map for coffee in Colombia; types describe agro-climatic zones, light yellow cells share aspects of multiple types, and grey area has a low suitability score.*

### *3.1.1.2 Municipalities with relevant coffee area*

Of the 1122 municipalities in Colombia 592 reported harvested coffee area. From these we selected 379 which cumulatively contained 95% of the total harvested area. The data was obtained from Marco Geoestadistico Nacional (MGN), on the Geoportal of DANE (ver https://bit.ly/2XLP9hj).

### *3.1.1.3 Map of forest cover*

The map of forest cover by IDEAM resulted from the Sistema de Monitoreo de Bosque y Carbono for the year 2018. In Colombia, forest land cover includes shrub land, palm trees, Guadua (bamboo), and grassland as long as tree cover dominates, the canopy density exceeds 30%, tree height is at least 5m and the area is at least of 1ha (IDEAM 2014).

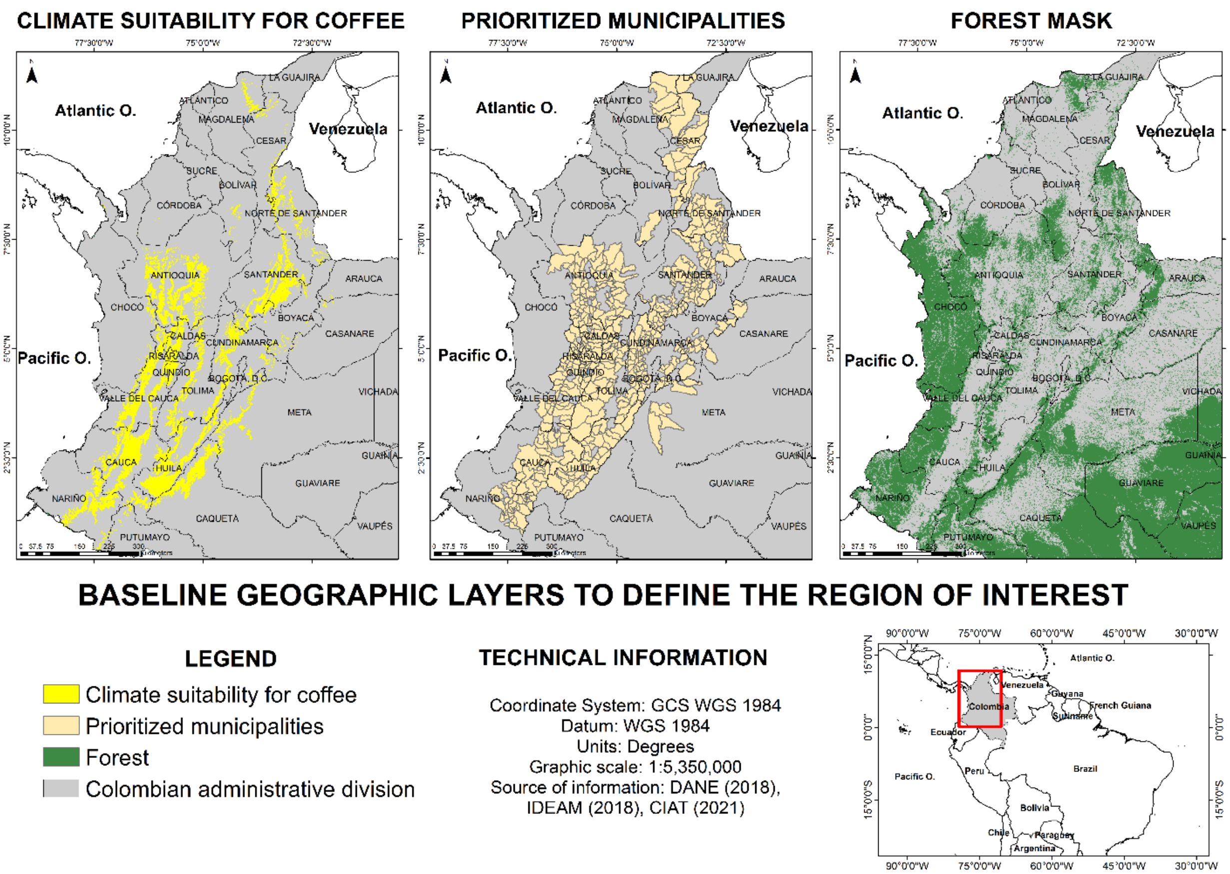



*Figure 4 Baseline layers to define the region of interest; A Climatic suitability for coffee, B Municipalities with coffee area, C Forest Cover.*

### *3.1.2 Satellite imagery*

We used satellite imagery from the collection S2 of the European Space Agency (ESA). The collection S2 at level 2A is based on the surface reflection, which contains bands of the optic spectrum, infrared and quality bands (compare https://bit.ly/3e49VCn), and a cloud probability band (compare https://bit.ly/380XvHp) which can be used to eliminate atmospheric errors. The temporality of the imagery is five days, which allows a continuous monitoring of the earth land cover. (Google Earth Engine n.d.). Additionally, we used Synthetic Aperture Radar (SAR) images from Sentinel-1 at a resolution of 10m. The selected images were taken from the period 2018-01-01 until 2018-12-31 and masked by the region of interest. In addition, we complemented the spectral information with elevation data from the Shuttle Radar Topography Mission (SRTM) with 30m resolution.

### *3.1.3 Training data*

We assembled a raw dataset of coffee georeferences from several sources. Raw data was obtained through previous projects including questionnaires to coffee farmers, the GBIF repository, from coffee industry and NGO partners and farmer training locations, and from local and national farmer organizations. The raw data was cleaned to exclude misplaced references, references with insufficient accuracy and duplicates.

With these raw data as a reference, a curated training data set was created. The manually curated training data from these polygons is the most valuable training data set to train the Random Forest classifier due to its precision.

In Google Earth pro polygons containing coffee, or other land uses (not coffee) were created where high resolution imagery was available for the years 2017 and 2018 within the region of interest. Figure 6 shows some of the polygons drawn in Google Earth Pro, where yellow is interpreted as coffee, and red polygons no-coffee.


*Figure 5 Polygons of coffee (yellow) and no-coffee (red)*

From within these polygons, we created 15,000 random training samples for the coffee class, and 30,000 for the non-coffee land uses. We set a minimum distance of 10m between samples (Figure 7).

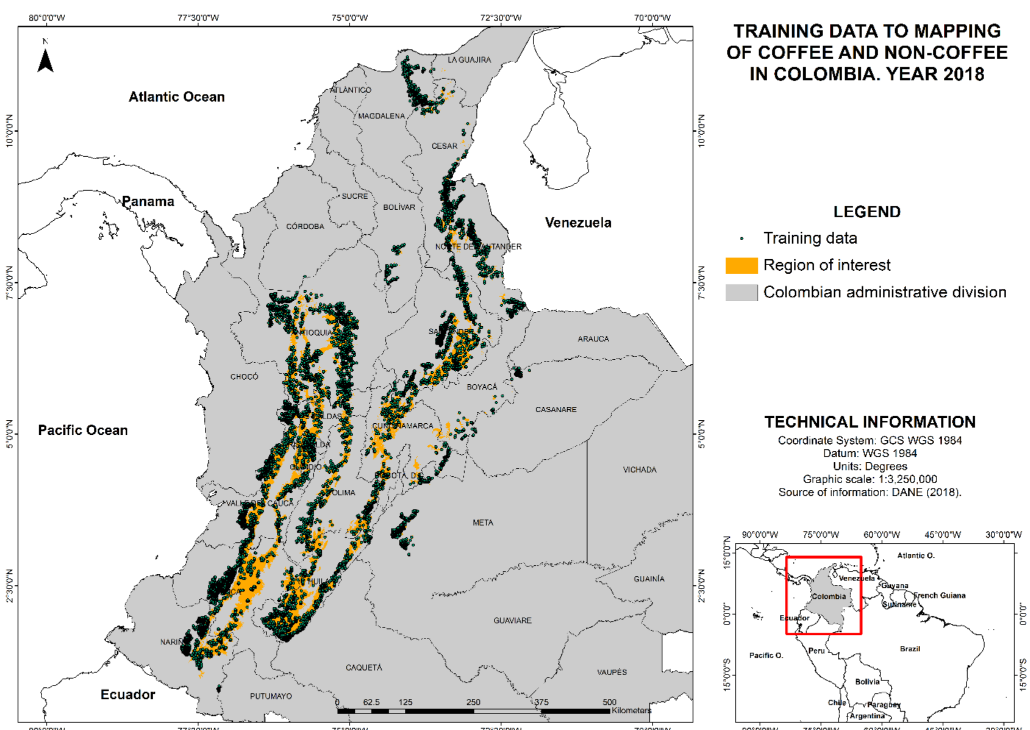



*Figure 6 Training data and region of interest*

### *3.1.4 Pre-processing of training data*

We further cleaned the initial training data using unsupervised clustering. Contaminated training data can reduce the accuracy of classification. Despite the manual generation of the data, the coffee polygons may include other land cover like farm houses or adjacent forest. The non-coffee polygons may include some coffee, or can confuse training by showing a diffuse 'all-other' signature. With the k-means clustering on Euclidean distances of input variables, the training data gets sorted into distinct groups with unique and separable spectral signatures.

We determined the optimal number of clusters using two methods: the Elbow method in the algorithm fviz_nbclust and the approach NBclust which generates an ensemble vote from 30 indices in Rstudio. According the Elbow method, the optimal number of clusters would be 4, while the ensemble vote indicated 2-4 (Figure 8).

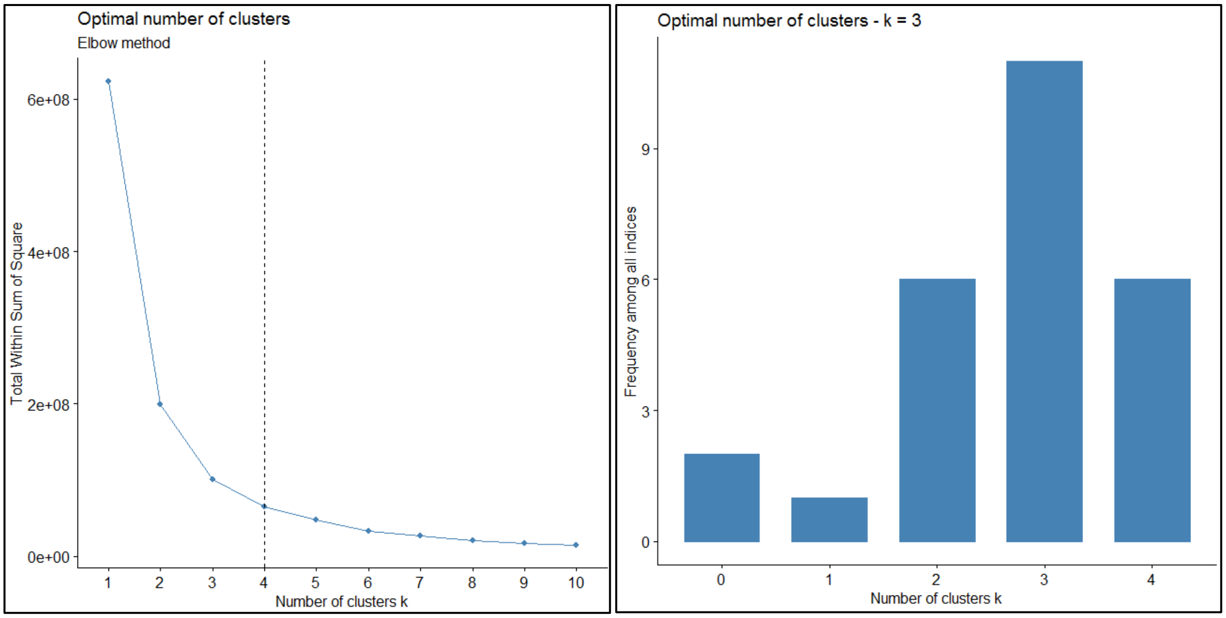



*Figure 7 Selection of clusters using the Elbow method (left), and the ensemble method (right)*

We determined that 4 clusters are optimal after further visual inspection of these clusters. Category 0 included predominantly grass land, field crops or shrub land. Category 1 corresponded to coffee, category 2 comprised forest cover and secondary vegetation, while category 3 included urban areas, bare soils, or water bodies. We checked the separability of these groups with density distribution plots of each variables. The plots showed that the signatures of coffee (cat 1) and forest cover (cat 2) are highly similar, while urban areas (cat 3) and grasslands (cat 0) are clearly distinct (Annex I).

## 3.2 Pre-processing of satellite imagery

All optical satellite imagery was pre-processed by applying the region of interest mask, and to reduce the information in the raw dataset of 5000+ available images for 2018.

### *3.2.1 Cloud cleaning*

The satellite images S2-2A were corrected using the product "Cloud Probability" to detect the probability of a pixel being a cloud. The values of this product range from 0 to 100%, where 0 represents a pixel of high quality and 100 corresponds to pixels with clouds. For the region of interest, we selected the pixels with a maximum cloud probability of 30% to include only images without atmospheric errors. Next, we selected the bands B2, B3, B4, B8, B11 and B12 which correspond to Blue, Green, Red, Near Infrared, Short Wave Infrared 1 and Short Wave Infrared 2. To these bands we applied the following steps:

- Cloud probability masking – this filter was applied to every image of the collection, masking and removing all pixels with a cloud probability above 30%.

- Cloud edge masking – in addition to cloud probability masking we removed the cloud edges, which weren't removed before, because for the 10m bands, these often aren't recognized. For this reason, the bands Edge red 4 and water vapor were used.

### 3.2.2 Vegetation index NDVI

We calculated the normalized difference vegetation index (NDVI) to reinforce the classification and achieve a differentiation between coffee and other land uses. We scaled each pixel by multiplying with 0.0001 to get an adequate range of values. Next, we used equation 1 to calculate NDVI (Schultz, et al., 2016).

$$NDVI = \frac{Near\ infrared - Red}{Near\ infrared + Red} \qquad \text{Eq. 1}$$

where Near Infrared corresponds to band B8 and Red to band B4.

### 3.2.3 Sentinel-1 image correction

The images available in GEE contain the data S1 SAR pre-processed with the toolbox for Sentinel-1 to generate a calibrated and ortho-corrected product. We followed steps to derive the coefficient of retro dispersion for each pixel, starting with the application of the orbita file, followed by the edge noise elimination GRD, thermal noise reduction, and radiometric calibration, and finally terrain correction using el 30m DEM SRTM or DEM ASTER (compare https://bit.ly/385Ujud).

### 3.2.4 GLCM texture indices

We calculate the co-occurrence matrix of grey levels using the tool glcmTexture in GEE which derives 14 metrics proposed by Haralick, et al., (1973) and 4 additional metrics by Conners et al (1984). Of the 18 bands produced by the tool, we selected entropy and dissimilarity for the polarizations VH and VV. In total we added 4 bands to the classifications model, were VH_dissimilarity, VV_dissimilarity, VH_entropy, and VV_entropy.

### 3.2.5 Reduction of imagery

After area exclusion and quality filter, we created a consolidated image from all available imagery for the year 2018, consisting of the spectral bands blue, green, red, swir1 and awir2. This 2018 image consists of statistical indicators for each band, including the minimum, maximum, mean and median, and standard deviation. Additionally, the digital elevation data was added, so that the final image consisted of 41 spectral bands.

*Table 1 Spectral bands in the final satellite image consisting of statistical indicators*

| Name of spectral band | Number of bands | | | | | | |
|---|---|---|---|---|---|---|---|
| | Digital level | Mean | Median | Standard deviation | Minimum | Maximum | Total |
| ***Blue*** | 0 | 1 | 1 | 1 | 1 | 1 | 5 |
| ***Green*** | 0 | 1 | 1 | 1 | 1 | 1 | 5 |
| ***Red*** | 0 | 1 | 1 | 1 | 1 | 1 | 5 |
| ***Nir*** | 0 | 1 | 1 | 1 | 1 | 1 | 5 |
| ***Swir1*** | 0 | 1 | 1 | 1 | 1 | 1 | 5 |
| ***Swir2*** | 0 | 1 | 1 | 1 | 1 | 1 | 5 |
| ***NDVI*** | 0 | 1 | 1 | 1 | 1 | 1 | 5 |
| ***Slope*** | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| ***Elevation*** | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| ***VH_entropy*** | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| ***VV_entropy*** | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| ***VH_dissimilarity*** | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| ***VH_dissimilarity*** | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| **Total bands** | | | | | | | 41 |

## 3.3 Classification of land cover and model validation

The land use classification was implemented in the platform RStudio, using the library RGEE to communicate with Earth Engine. This approach enabled the optimization of RandomForest parameters with the R libraries mlr3, RandomForest and RFE. The process consists of the following steps:

1. Read files in GEE: all spatial data is stored in the asset of GEE, such as the pre-processed Sentinel-2A and Sentinel-1 data for 2018 from the previous step, and the training data.
2. Definition of objective: For all training data the classification objective is specified by adding the respective value to each reference location. The values were either coffee, or one of the non- coffee classes, represented by values 0 - 3 respectively.
3. Separation of training data: The data set is split 70/30% into a training and validation dataset.
4. Extraction of data: For each training and testing location the data for each spectral bands is extracted from the imagery.
5. Variable selection: introducing variables which are correlated can result in overfitting of the classifier. We therefore used the libraries Boruta and RFE to select variables. These tools select the variables which most contribute to the classification model, based on the variables importance attribute in RandomForest. Both algorithms suggested to use all 41 variables (Annex 2).
6. Cross-validation and tuning: The training data is divided into 5 groups for cross-validation. Different combinations of key tuning parameters for Random Forests (Node size, variables picked (mtry), and number of trees - Table 2) are used and the accuracy on each data subset is evaluated (mlr library):

*Table 2 Parameter spaces tested for Random Forest tuning*

| Number of trees | | mTRY | | Node number | |
|---|---|---|---|---|---|
| Minimum | Maximum | Minimum | Maximum | Minimum | Maximum |
| 250 | 1000 | 2 | 41 | 2 | 41 |

The parameter combination which performed highest according to the accuracy metric was used to train the classification algorithm:

- Mtry = 17
- Node = 2
- Ntree = 500

7. RandomForest classification: Finally, we used the Random Forest classifier trained in this process to classify the pre-processed summary image in Google Earth Engine. The classifier was tested on the internal cross validation data (Table 3).

*Table 3 Cross validation classification result*

| | Grassland | Coffee | Forest | Urban areas | Class error |
|---|---|---|---|---|---|
| **Grass land** | 4408 | 95 | 58 | 12 | 0.036 |
| **Coffee** | 202 | 4050 | 283 | 1 | 0.107 |
| **Forest** | 99 | 397 | 3820 | 2 | 0.115 |
| **Urban areas** | 39 | 7 | 4 | 4448 | 0.011 |

### 3.4 Post-processing of the classification

Pixel based algorithms, as used here, often produce a “salt and pepper effect”, because each pixel is evaluated in isolation. We applied post procession to the results to reduce the impact of this effect and better generalize the map. Five spatial processing steps were used: reclassification, majority filter, region groups, cell elimination, and gap filling.

The first step was to reclassify classes which aren’t of interest into a single “no-coffee” class. We merged grid cells from the classes “grassland”, “forest” and “urban” into a “no coffee” class (Figure 9).

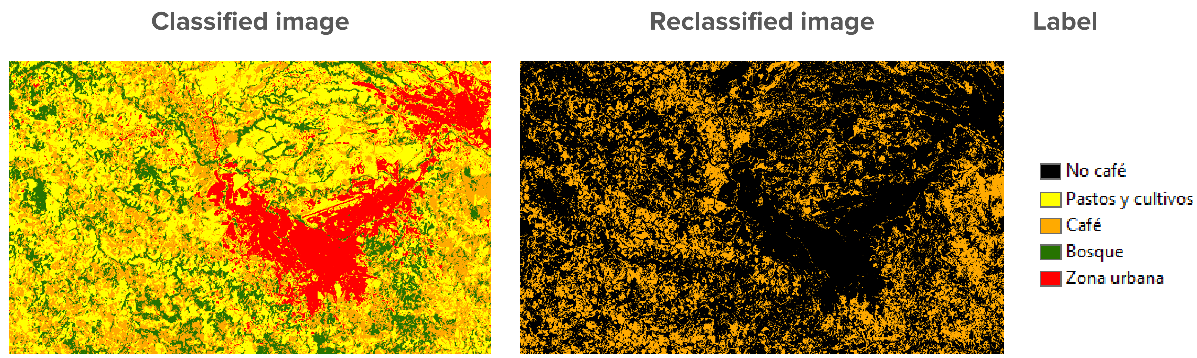



*Figure 8 Unprocessed classification result and reclassified image with only coffee/no-coffee*

We then applied the function ‘majority filter’. The function replaces cells in a raster based on the majority of their contiguous neighboring cells. The neighborhood consists of a window of 9 pixels and the central pixel will change its value if the majority of its neighbors have the same value. This is illustrated in Figure 10 (Esri, 2016).

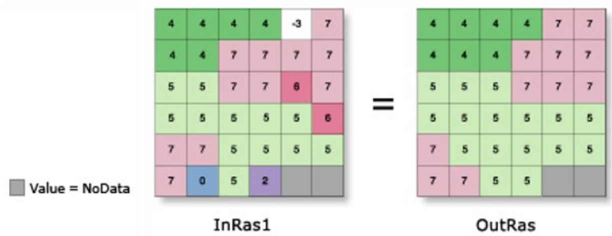



*Figure 9 Majority filter tool illustration (ESRI 2016).*

Next, the region groups filter assigns a unique number to each region. Regions are a contiguous set of cells of the same zone type. When the regions need to be processed separately, each must be identified as a separate entity. The Region Group tool assigns a new value to each region in a raster as shown in Figure 11 (Esri, 2016)

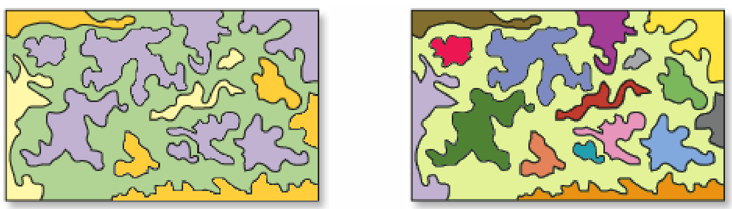


*Figure 10 Region groups. Left: regions separated by a single value, Right: Disconnected regions are assigned a unique value (ESRI 2016).*

To reduce spurious grid cells, we assigned NoData values at region groups equivalent to 0.05ha to 1ha. The tool SetNull returns NoData if this conditional evaluation is true, and returns the value specified by the original raster if it is false (Figure 12).

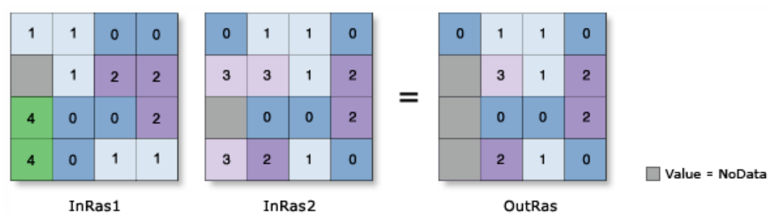



*Figure 11 Elimination of spurious groups using SetNull in Arcgis (ESRI 2016)*

Finally, the information for these grid cell groups was filled with the values from neighboring fields with the tool Nibble from ArcGis. Nibble replaces cells of a raster corresponding to a mask with the values of the nearest neighbors (Figure 13).

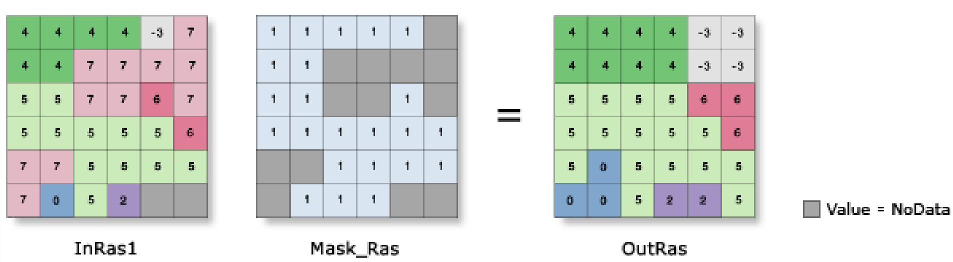



*Figure 12 Gap filling using the tool Nibble in ArcGis (ESRI 2016).*

The generalization of the initial results map of Coffee/no-coffee áreas varied depending on the choice of minimum area filter size (Figure 14). With a mínimum area of 0.05ha some riparian forest, which escaped the IDEAM forest cover maps, was shown as “Coffee”. A filter size 1ha led to a removal of areas which in the base map appear to be coffee patches.

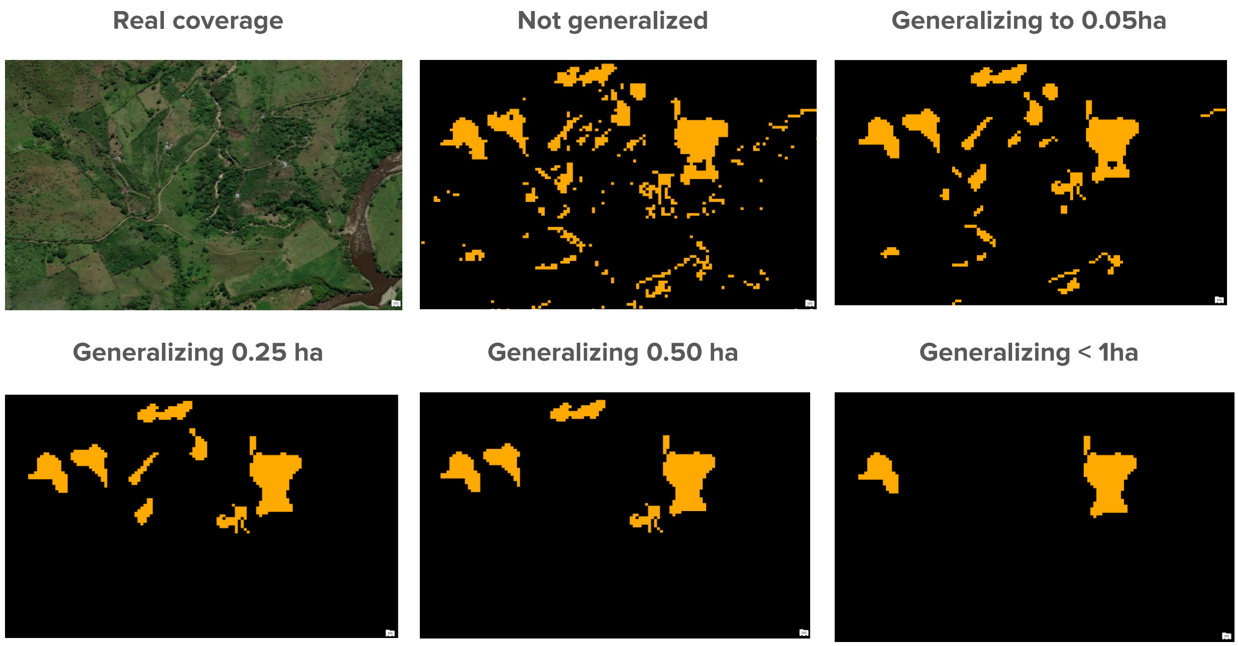



*Figure 13 Generalization effect of different majority filter choices*

## 3.5 Validation

We validated the result map by comparing reported area in national statistics by municipality and modelled area, and second by creating a reference pixel database on high resolution satellite data. We identified regions of low classification performance and repeated the steps in “3.1.3 - Training data” and the following twice.

### *3.5.1 Comparison of reported and mapped area*

We compared the modeled area according to the classification into coffee/non-coffee, with reported planted area statistics of coffee for the year 2018 to identify regions with low model accuracy. Area statistics were reported at municipal level, so that we estimated modeled area for each municipality. We then calculated the difference between the two data sources. We differentiated municipalities with underestimation of area (0 – 100% of reported area), medium error (100 – 250% of reported area) and high error (>250% of area).

The resulting map was evaluated to prioritize regions for additional training data sampling (Figure 11). We further summarized it by counting the number of municipalities with high error rates (Table 3). Following the visual inspection of the map, and review of the table, we prioritized Antioquia, Cundinamarca, Santander, and Cauca for additional sampling. With the added training data, we repeated the classification steps. After the first repetition, the map of coffee/non-coffee had 21 municipalities less in the “high error” category (Figure 15 and Table 3).

*Table 4 Number of municipalities with high classification error between initial and final model*

| Department | Initial model | Preliminary model | Corrected |
|---|---|---|---|
| | Number of municipalities | Number of municipalities | municipalities |
| Antioquia | 33 | 10 | 23 |
| Boyacá | 15 | 03 | 12 |
| Caldas | 03 | 01 | 02 |
| Cauca | 20 | 06 | 14 |
| Cesar | 01 | 01 | 00 |
| Cundinamarca | 29 | 15 | 14 |
| Huila | 03 | 01 | 02 |
| Magdalena | 01 | 01 | 00 |
| Meta | 03 | 00 | 03 |
| Nariño | 03 | 00 | 03 |
| Norte de Santander | 07 | 06 | 01 |
| Quindío | 05 | 01 | 04 |
| Risaralda | 02 | 01 | 01 |
| Santander | 27 | 06 | 21 |
| Tolima | 02 | 00 | 02 |
| Valle del Cauca | 18 | 04 | 14 |

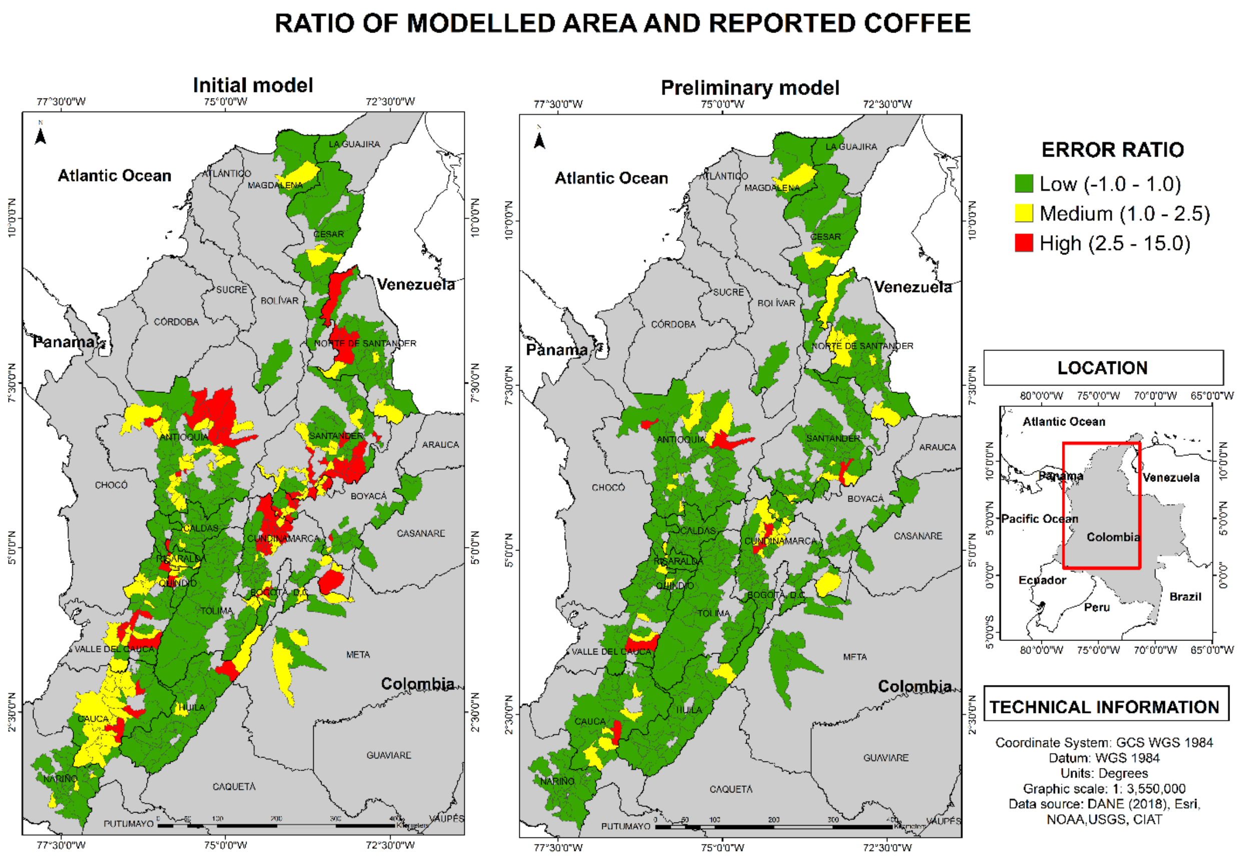



*Figure 14 Ratio of modelled area and reported coffee area.*

To evaluate the final map, we differentiated 5 error classes from very low to very high. Most municipalities now showed an acceptable error (0-300% of reported area). In the departments Antioquia, Cundinamarca, and Santander some municipalities still showed high error (Table 5).

*Table 5 Number of municipalities by error class in each department*

| DEPARTAMENT | NUMBER OF MUNICIPALITIES IN EACH ERROR CLASS | | | | |
|---|---|---|---|---|---|
| | Very low | Low | Medium | High | Very high |
| **ANTIOQUIA** | 38 | 28 | 7 | 3 | 0 |
| **BOLIVAR** | 1 | 0 | 0 | 0 | 0 |
| **BOYACÁ** | 19 | 10 | 3 | 1 | 0 |
| **CALDAS** | 21 | 2 | 0 | 0 | 0 |
| **CASANARE** | 3 | 0 | 0 | 0 | 0 |
| **CAUCA** | 12 | 17 | 1 | 2 | 0 |
| **CESAR** | 12 | 1 | 0 | 0 | 0 |
| **CUNDINAMARCA** | 42 | 13 | 4 | 4 | 2 |
| **HUILA** | 30 | 4 | 0 | 0 | 0 |
| **LA GUAJIRA** | 1 | 1 | 0 | 0 | 0 |
| **MAGDALENA** | 3 | 0 | 0 | 0 | 0 |
| **META** | 7 | 0 | 0 | 0 | 0 |
| **NARIÑO** | 26 | 7 | 0 | 0 | 0 |
| **NORTE DE SANTANDER** | 24 | 6 | 0 | 0 | 0 |
| **QUINDIO** | 5 | 4 | 2 | 0 | 0 |
| **RISARALDA** | 11 | 2 | 0 | 0 | 0 |
| **SANTANDER** | 37 | 12 | 5 | 2 | 1 |
| **TOLIMA** | 34 | 0 | 0 | 0 | 0 |
| **VALLE DEL CAUCA** | 23 | 12 | 1 | 1 | 0 |

Visual inspection of the map showed that distribution of error no longer showed a clustered spatial distribution (Figure 16).

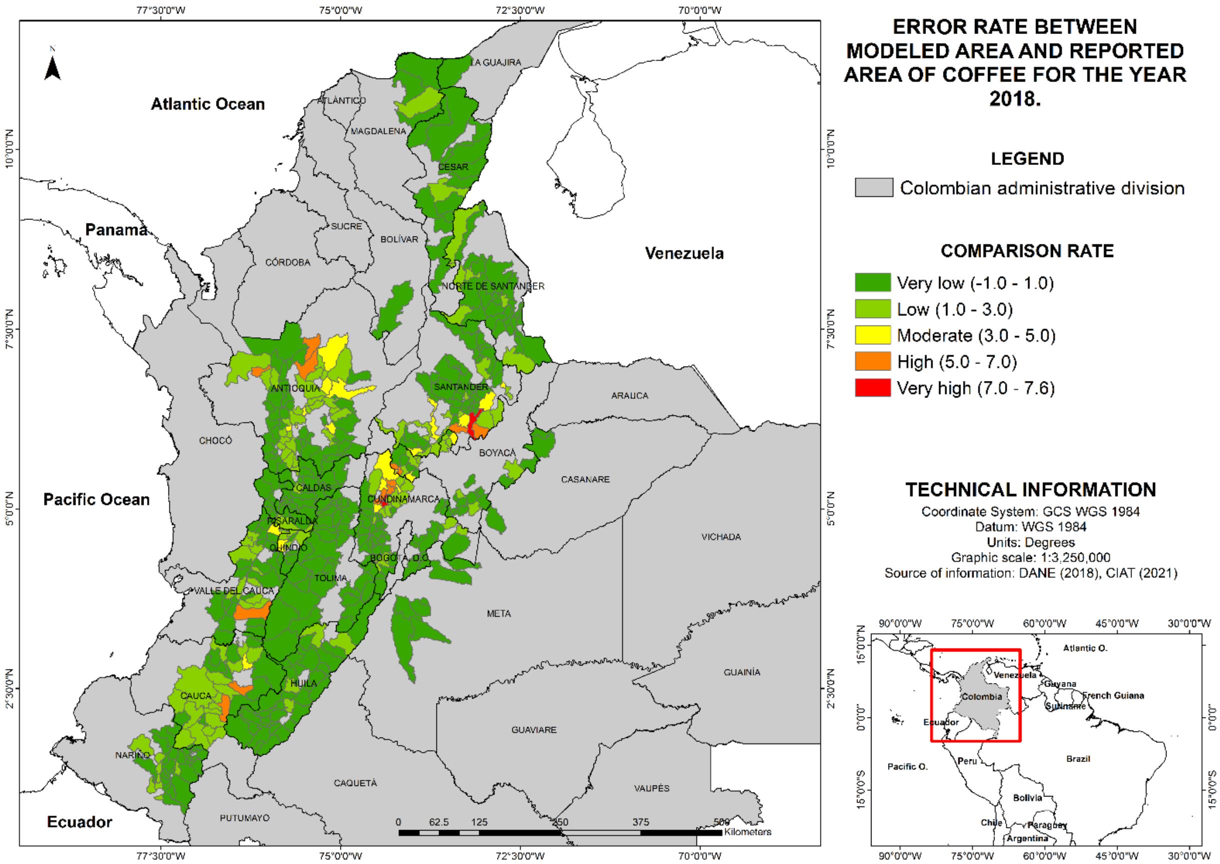



*Figure 15 Error rate after second targeted sampling by municipality. Error is shown from very low (dark green) to very high (red).*

### *3.5.2 Reference data from high-resolution satellite data*

We used stratified sampling of reference data from high resolution satellite imagery to validate classification accuracy following methods suggested by Olofsson, (2014). We used the SEPAL tool (https://sepal.io) to generate a random stratified location sample for validation. The tool accounts for the expected prevalence of land use classes, and the classification result of the preliminary map. The tool creates samples from the entire region of interest for each land use class. In addition to prevalence, sampling depends on the expected accuracy and desired confidence intervals. In our case, the coffee class is considered to be a relatively rare class, compared to other land uses in the region of interest. For this reason, we defined the expected accuracy to be 60% and for non-coffee land use 90%. The tool generated 1030 locations for validation, of which 400 were for coffee, and 630 for non-coffee. The spatial distribution can be seen in the following map (Figure 12)

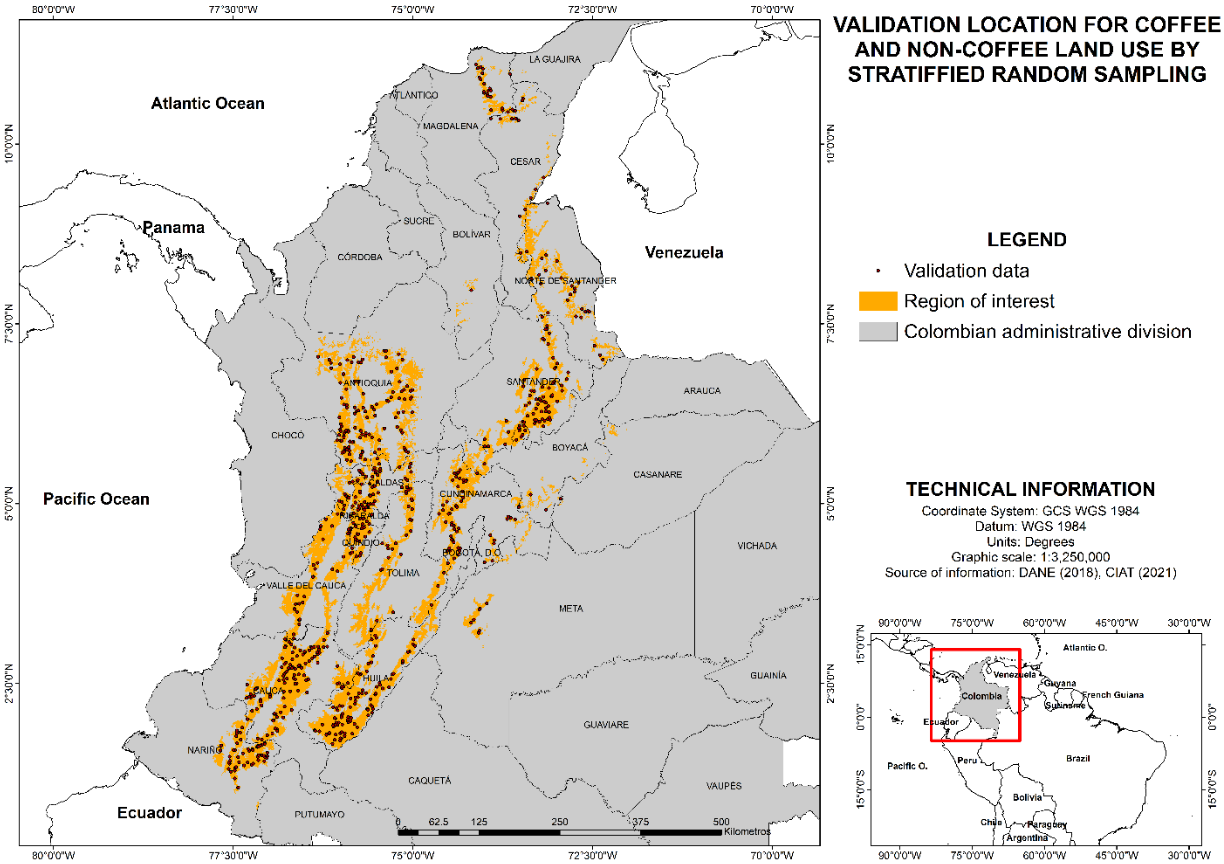



*Figure 16 Validation locations for coffee and non-coffee land use by stratified random sampling.*

For 745 references of the 1030 locations satellite imagery of sufficient quality was available to evaluate with confidence whether a location was coffee or non-coffee. The coffee region in Colombia has a high cloud cover which limited the availability of high resolution data.

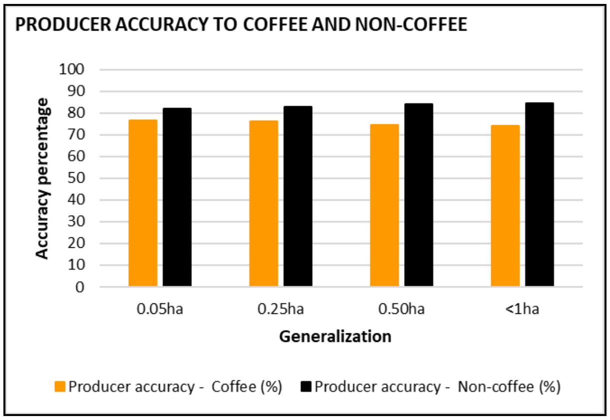
Using these data, we created confusion matrices for each generalization from 0.05 to 1ha (Annex III). After visual examination of the generalized maps, and the confusion matrix results, we decided that 0.5ha is the most adequate minimum plot size filter.

## 4 Results

The first step of the map quality assessment was to generate a set of confusion matrices for each step of the generalization process. From these matrices, we extracted a set of indicators including, the producer and user accuracy for coffee and non-coffee classes as well as the overall accuracy

The results show that the producer accuracy for each of the generalizations step ranges between 70% and 80% for the coffee class, which is equivalent to omission errors between 20% and 30%. On the other hand, the non-coffee producer accuracy was significantly higher values ranging between 80% and 90%. Coffee being a relatively rare class in comparison with the non-coffee one, it was expected to observe a greater probability of a false negative.

The user accuracy represents the probability that an area predicted to be of a certain class really is that class. Our results show that the highest user accuracy was obtained for a generalization of 0.5ha with a value greater than 70% for coffee. For this reason and for the fact that there are coffee plantations of less than a hectare in Colombia, as shown in in section 7.2, we selected the generalization of 0.5ha for our final results.



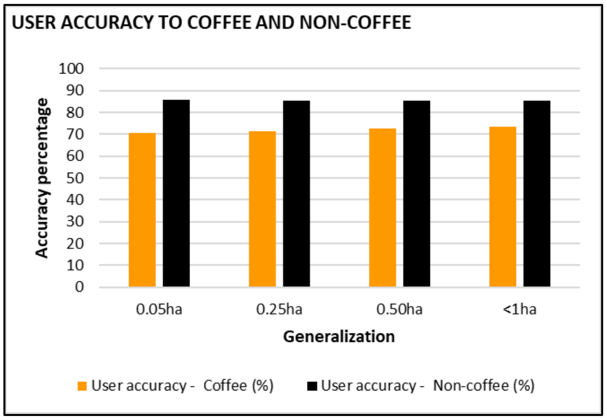




*Figure 17 Producer and user accuracy for different levels of generalization*

With the final coffee map and taking into account the producer and user accuracies, we estimate that there are approximately 1'093'249ha of coffee ± 136. 381ha in Colombia. The departments identified with the largest coverage of coffee plantation are Huila, Tolima, Antioquia, and Cauca, with 16.9%, 13.2%, 13.1%, and 11.1% of their land used for coffee production respectively. On the other hand, the departments with the lowest percentages are Meta (0.36%), Casanare (0.30%), Bolivar (0.15%) and La Guajira (0.02%).

Based on the EVA data, we produced the error rate by department (Figure 20). 13 out of 19 departments are showing that our results are over-estimating coffee coverage by an average factor of 0.65. On the other hand, 6 of 19 departments showed a potential underestimation of coffee area with an average factor of -0.54.

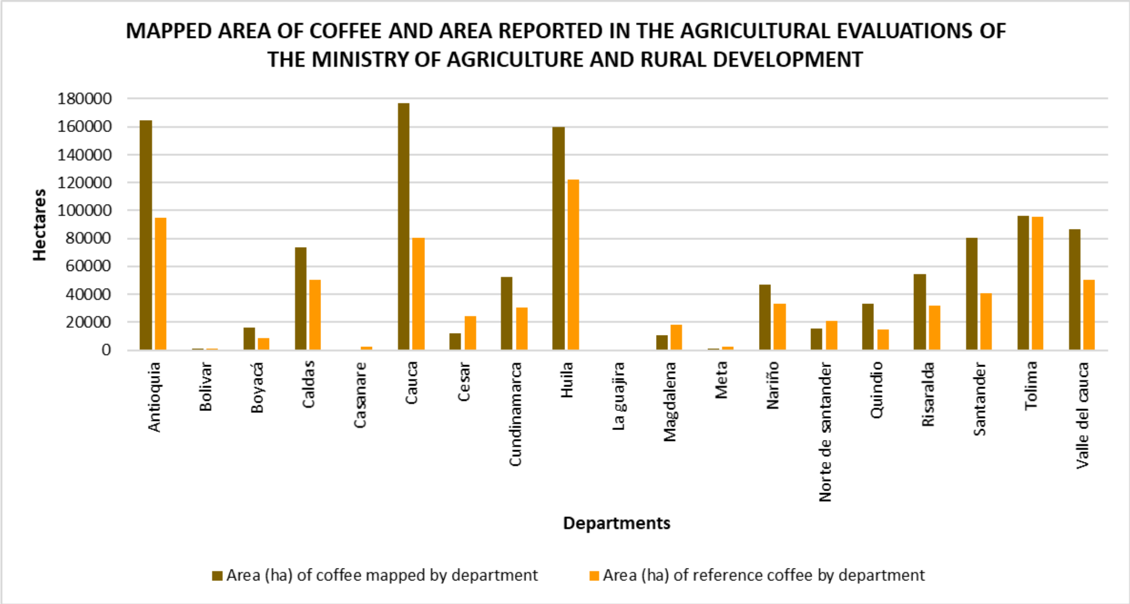



*Figure 18 Reported (EVA - MARD) and modelled area by department.*

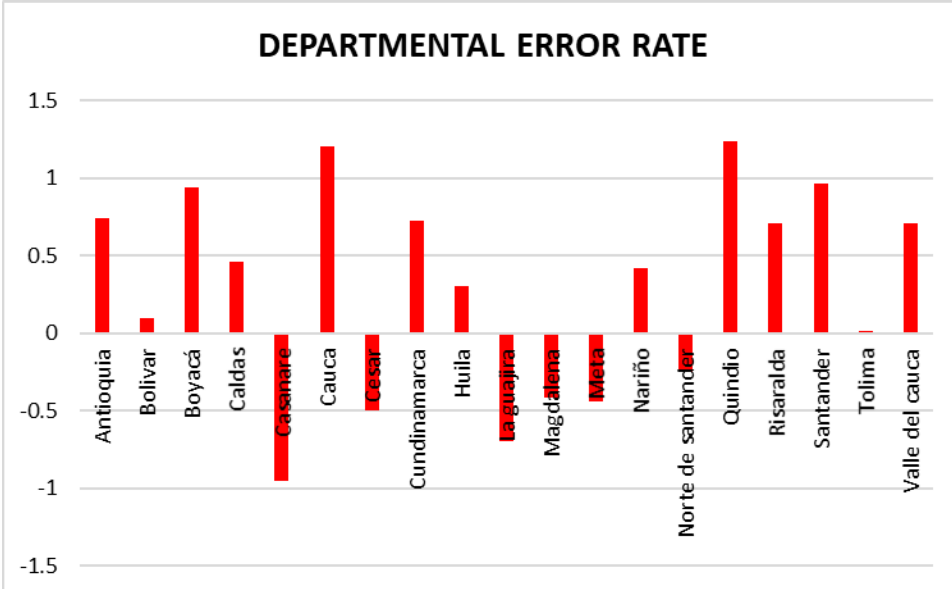



*Figure 19 Error rate by department*

Finally, Figure 21 shows the spatial distribution of coffee and non-coffee area in Colombia for 2018, with an overall accuracy of 80.6%.

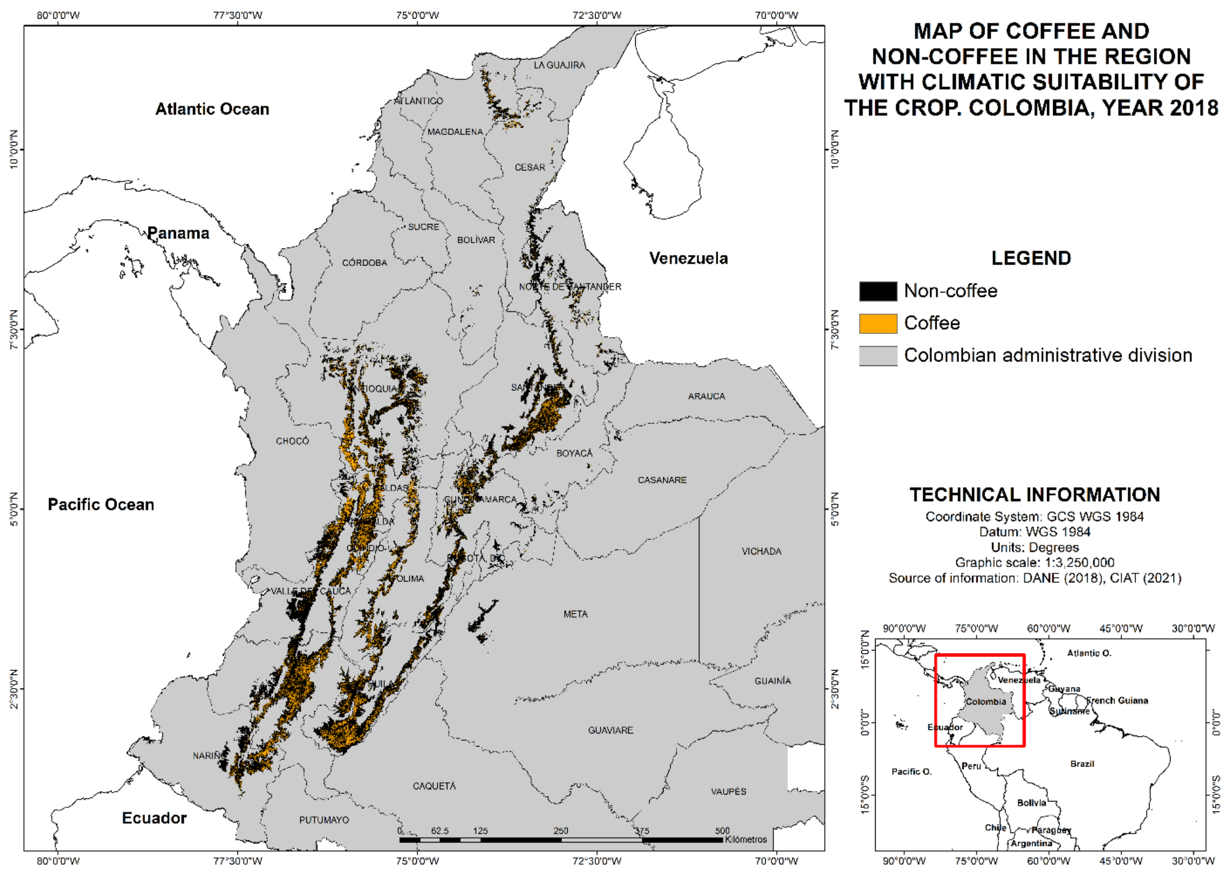



*Figure 20 Final coffee map for Colombia for the year 2018.*

## 5 Discussion

Knowledge about the distribution of coffee production is a valuable resource to improve data driven decisions support. We demonstrated how precise maps of coffee production can be produced at country scale from freely available remote sensed data. The resulting overall accuracy of 81%, and the user accuracies of 72% on the coffee class, and 85% on the non-coffee class, are in line with previously reported accuracies (Hunt et al. 2020). However, previously, mapping was either restricted to areas where coffee production can be expected to be a dominant land use class. A key innovation of our work is the use of climate suitability data to delineate the region of interest, making a whole-country mapping computationally feasible. Recent country scale mapping for coffee in Vietnam (Maskell et al. 2021) and cocoa in Ghana and Ivory coast (Abu et al. 2021) showed lower user accuracies, despite using similar data and methods.

Coffee being a relatively rare class in comparison with the non-coffee one, it was expected to observe a greater probability of a false positive (non-coffee area being identified as coffee). This is leading to an overestimation of coffee areas in 13 out of 19 departments when compared to the Agricultural Evaluation from the Colombian Ministry of Agriculture and Rural Development. The results however stay within acceptable limits with a producer accuracy of 80% for the coffee class and a user accuracy above 90% for the non-coffee class.

A key challenge was the generation of sufficient training data. Available data sources were not collected for the purpose of informing remote sensing. Geo-references of coffee commonly describe supply chains and may include farm houses, primary processing facilities or collection stations. Such data is valuable to describe flows of coffee and to identify broad origins, but cannot be used to train remote sensing. Even plot level data, collected to estimate coffee area, remotely observe pest and disease outbreaks or similar uses was not fit for this purpose as adjacent forest may be included, riparian areas or minor patches of other land cover. We therefore had to invest substantial effort into generating clean data from such initial reference locations. Despite the expert generation of polygons, unsupervised classification of the input data revealed the false inclusion of pixels with diffuse land cover. Especially forest cover showed a spectral signature very similar to the coffee signature.

In this work, we developed a novel approach to reduce the scope of the mapping area based on coffee suitability maps and the inclusion of environmental variables in the analysis. This approach has been shown to be very effective to serve as a first filter to pinpoint the areas the need a precise mapping and discard large areas where coffee is very unlikely to be cultivated, thus saving a large amount of processing time.

Our mapping is not the first attempt at whole country mapping of coffee areas, but likely the most accurate approach available to date. Previous work was largely reliant on subnational production statistics, sometimes in combination with environmental or satellite data (Monfreda, Ramankutty, and Foley 2008; You et al. 2014). Such approaches were sufficiently accurate to inform global analysis, but not accurate enough for decision making by stakeholders (Bebber, Castillo, and Gurr 2016; Eriyagama, Chemin, and Alankara 2014).

Our data is similar in precision as other remote sensed data, such as weather observation, deforestation monitoring or general land use maps. Our mapping can therefore be useful to develop added value services. Coffee production can be beneficial for biodiversity and carbon stocks, if it spares primary forest, and is established in previously degraded landscapes (Martin et al. 2020).The combination of our coffee map and deforestation monitoring will make zero-deforestation commitments by the sector more feasible to implement. The combination with other land cover data can inform the identification of areas of potential coffee expansion. In addition, climate services for coffee will benefit from improved risk and productivity assessments.

To improve the results, very promising approaches, based on convolutional neural network, could be applied to better take into account the structure of the canopy. Indeed, the pixel-based analysis used in this study, Random Forest, does not use the full potential of high-resolution imagery such as Sentinel 1 and Sentinel 2. Future work should investigate more this type of deep-learning approach to refine the current map.

# INDONESIA COFFEE LAND USE MAP: METHODOLOGY

***Authors***
Jorge Andres Perez Escobar. Research associate. Alliance of Bioversity and CIAT.
Fabio Alexander Castro. Research associate. Alliance of Bioversity and CIAT.
Paula Andres Paz. Research associate. Alliance of Bioversity and CIAT.
Louis Reymondin. Scientist, Terra-i Project coordinator. Alliance of Bioversity and CIAT.
Christian Bunn. Scientist, Climate Action. Alliance of Bioversity and CIAT.

***Image classification team***
Jorge Andres Perez Escobar. Research associate. Alliance of Bioversity and CIAT.
Fabio Alexander Castro. Research associate. Alliance of Bioversity and CIAT.

***Recommended citation***
Perez, J.; Castro, F.; Paz, P.; Reymondin, L; Bunn, C. (2021). Mapping coffee land use in Indonesia for the year 2018. Alliance Bioversity & CIAT. Cali, Colombia.

## 1 Overview

This document provides a brief summary of the methods used to develop a map of coffee and non-coffee land uses for the year 2018 in Indonesia. The method consists of five general steps: initial training data collection, satellite data pre-processing, algorithm training, data classification and results post processing, and fourth, results validation. The initial step consists of the compilation and cleaning of reference data of the land use classes coffee and other land uses in the form of polygons in Google Earth Pro (GEP), and satellite data collection from Sentinel-2 Level 2A (surface reflectance), Sentinel-1 GRD, digital elevation model SRTM on Google Earth Engine (GEE). The satellite data is pre-processed to generate summary images for the year 2018. These data were used to train the RandomForest classifier to map coffee and non-coffee land uses. Spurious classifications were removed during post-processing. The final result was evaluated following Olofsson et al. (2014) and a reference data set created from high resolution images in GEP. Validation demonstrated an 85.4% overall accuracy and a user accuracy of 72.2 and 92.1% for coffee and non-coffee land use respectively. In the following, we will provide a more detailed overview of the methods and the results.

## 2 Data and methods

The methods used to create a map of coffee vs other land uses for Indonesia consists of four main steps as shown in Figure 2. The first step is the collection of input data, the second the pre-processing of satellite images, the third the land use classification and, last, validation of the model with evaluation data.

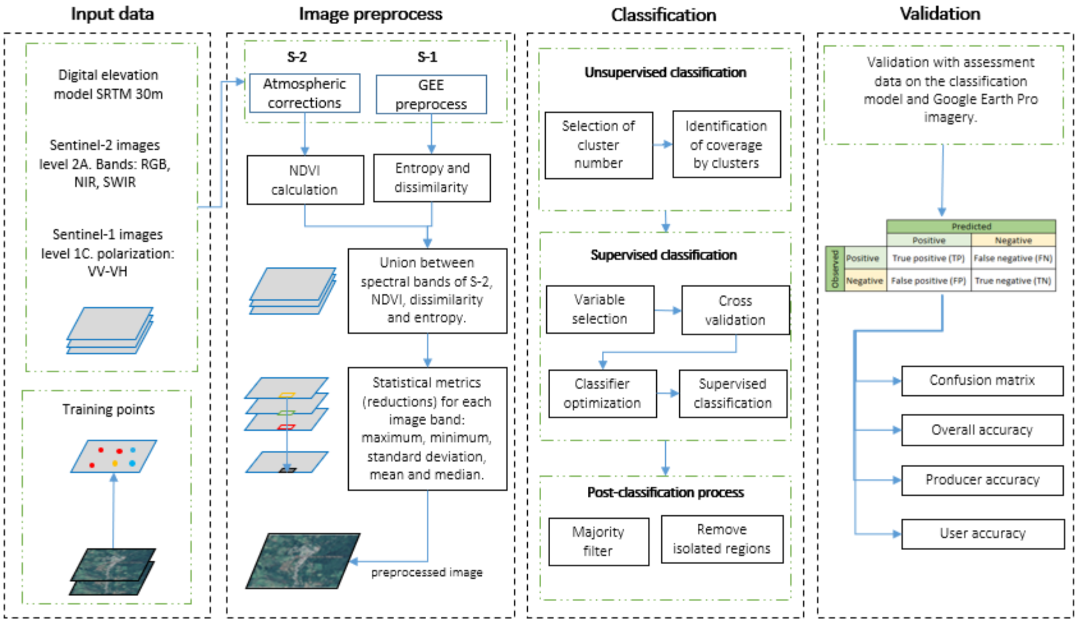



*Figure 21 Overview of methodology*

## 2.1 Input data

Three types of information were used for the land use classification into coffee and non-coffee in Indonesia. These were the creation of the zones of interest, satellite data, and training data. The first data set described the zones which were subject to classification into coffee and no-coffee. The second consists of the collection of high resolution multispectral satellite imagery fromSentinel-2A (S2) and Sentinel-1 (S1) with spatial resolution of 10m, and additionally the digital elevation model STRM of 30m, obtained through the data repository of Google Earth Engine (GEE). The third data set was created by collecting presence samples of coffee and non-coffee in the platform Google Earth Pro (GEP).

### *2.1.1 Region of interest*

The region of interest are areas which we can feasibly expect to contain coffee area. Exclusion of unfeasible regions reduces processing resources requirements and increases classification accuracy. This data was constructed at 1km spatial resolution by using three geographical input data: a map of the climatic suitability for coffee in Indonesia, the map of forest/no forest from GLAD for the year 2018, and lastly municipalities with relevant coffee area.

With this information, we delineated the region of interest by exclusion of a) pixels not climatically suitable for coffee, b) pixels with forest cover according to GLAD, and c) municipalities which cumulatively include 90% of all harvested area of coffee in Indonesia. The result can be seen in Figure 3.

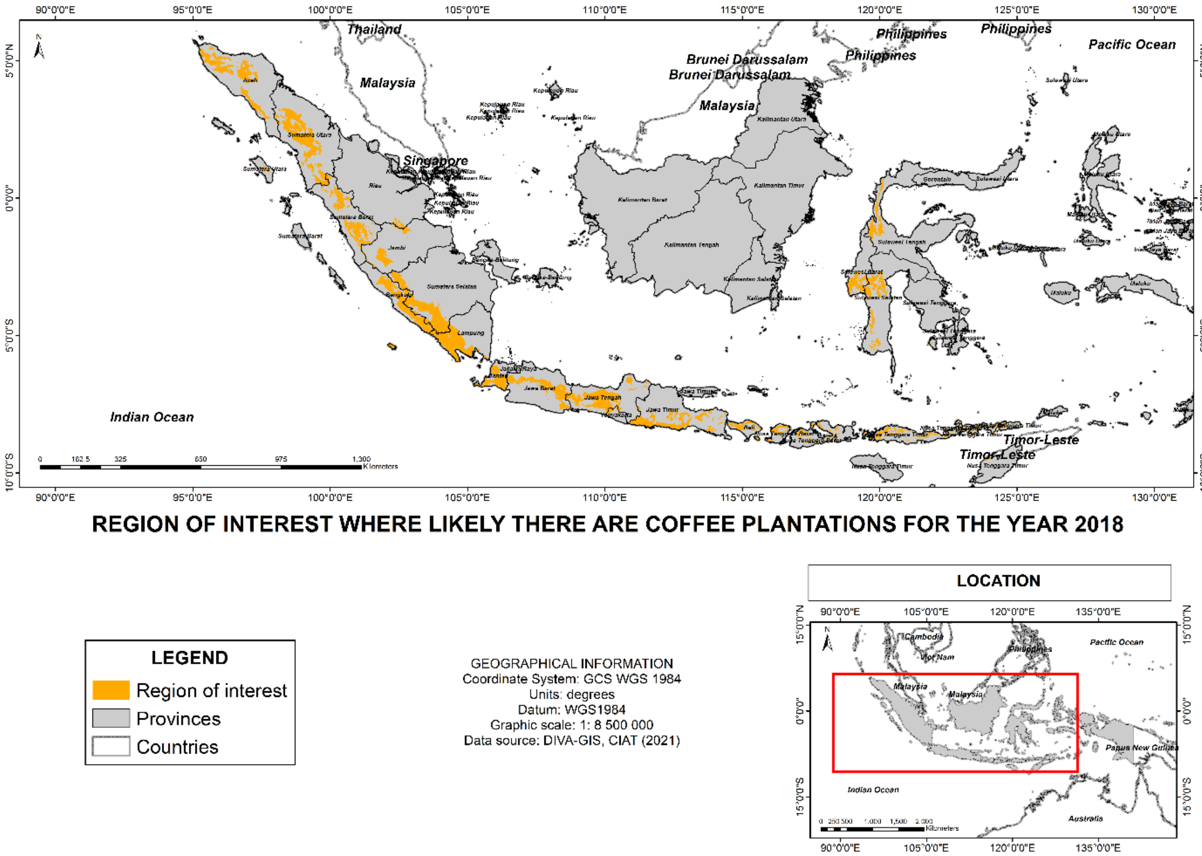



*Figure 22 Region of interest for classification; orange pixels can feasibly be expected to contain coffee area in 2018.*

In the following, we further describe the input data used for the zone of interest definition (Figure 5).

#### 2.1.1.1 Map of suitable area for coffee

The potential distribution of coffee production is strongly limited by climatic suitability. Modeling suitability with Random Forest has been shown to provide good results (Valavi et al. 2021). Therefore, we generated a climate suitability map for coffee in Indonesia by classifying climate data with a Random Forest classifier (Breiman 2001). The classifier was trained with the raw presence data set as described below, and WorldClim climate data at 1km resolution (Fick and Hijmans 2017). WorldClim provides monthly climate data for 1980-2010 (“current”) from which we generated 19 bioclimatic variables. The variables describe seasonality and extremes of temperature and precipitation throughout the year. The RF classifier was trained on this data and extrapolated on the climate maps. The result was a map which shows the similarity of a location with the climate at current coffee locations (Bunn et al. 2015). We assumed that locations with a suitability value lower than the 1st percentile at presence locations are climatically unfeasible for coffee production.

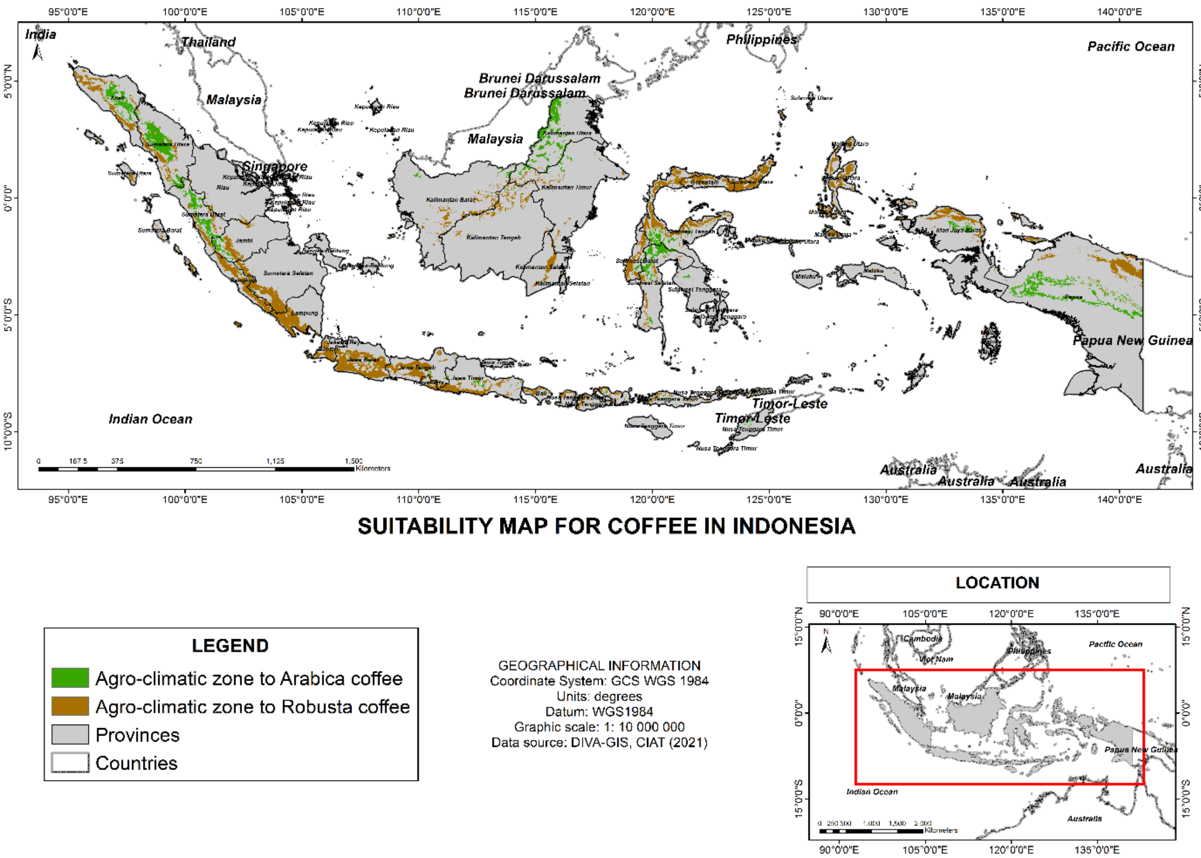



*Figure 23 Suitability map for coffee in Indonesia; types describe agro-climatic zones, light yellow cells share aspects of multiple types, and grey area has a low suitability score.*

### *2.1.1.2 Municipalities with relevant coffee area*

According to Directorate General of Estate Crops (DGEC 2018), coffee is produced in 32 of the 34 provinces, and 291 of 514 regencies, in Indonesia. We included in our analysis the provinces which together form continuous geographic units and include 95% of coffee area. From the regencies in these provinces, we selected the 184 regencies which cumulatively produce 95% of Arabica or Robusta in the provinces. Thus, our region of interest covered 90% of Indonesian coffee growing areas according to DGEC, while considering about 1/3 of regencies.

### *2.1.1.3 Map of forest cover*

The map of forest cover by Global Land Analysis Discovery (GLAD) includes data for Primary Humid Tropical Forests for the year 2018. This forest mask has a spatial resolution of 30 meters and is produced using Landsat images. We applied this mask to remove primary forest from the analysis

## BASELINE GEOGRAPHIC LAYER TO DEFINE THE RIGION OF INTEREST

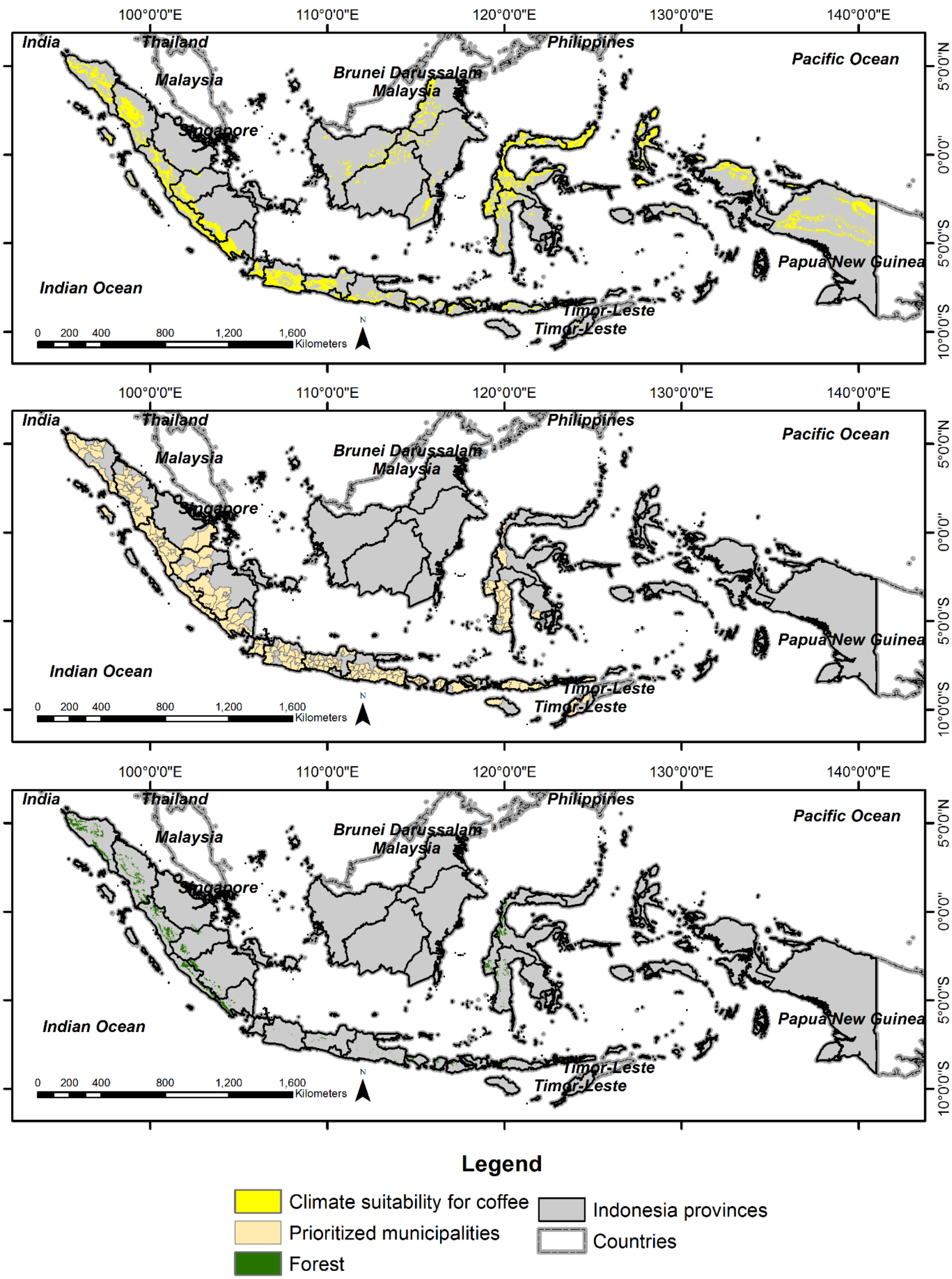


*Figure 24 Baseline layers to define the region of interest; A Climatic suitability for coffee, B Municipalities with coffee area, C Forest Cover.*

### *2.1.2 Satellite imagery*

We used satellite imagery from the Sentinel-2 collection of the European Space Agency (ESA). The collection Sentinel- 2A is based on the surface reflectance, which contains multi-spectral bands of the optic spectrum, infrared and quality bands (compare https://bit.ly/3e49VCn), and a cloud probability band (compare https://bit.ly/380XvHp) which can be used to eliminate atmospheric errors. The temporality of the imagery is five days, which allows a continuous monitoring of the earth land cover (Google Earth Engine n.d.). Additionally, we used Synthetic Aperture Radar (SAR) images from Sentinel-1 at a resolution of 10m. The selected images were taken from the period 2018-01-01 until 2018-12-31 and masked by the region of interest. We complemented the spectral information with elevation data from the Shuttle Radar Topography Mission (SRTM) with 30m resolution.

### *2.1.3 Training data*

We assembled a raw dataset of coffee georeferenced from several sources. Raw data was obtained through previous projects including questionnaires to coffee farmers, the GBIF repository, from coffee industry and NGO partners and farmer training locations, and from local and national farmer organizations. The raw data was cleaned to exclude misplaced references, references with insufficient accuracy and duplicates.

With these raw data as a reference, a curated training data set was created. The manually curated training data from these polygons is the most valuable training data set to train the Random Forest classifier due to its precision.

In Google Earth pro polygons containing coffee, or other land uses (not coffee) were created where high resolution imagery was available for the years 2017 and 2018 within the region of interest. Figure 6 shows some of the polygons drawn in Google Earth Pro, where yellow (pastures and crops) and purple (oil palm) polygons are interpreted as non-coffee , and brown polygons as coffee.

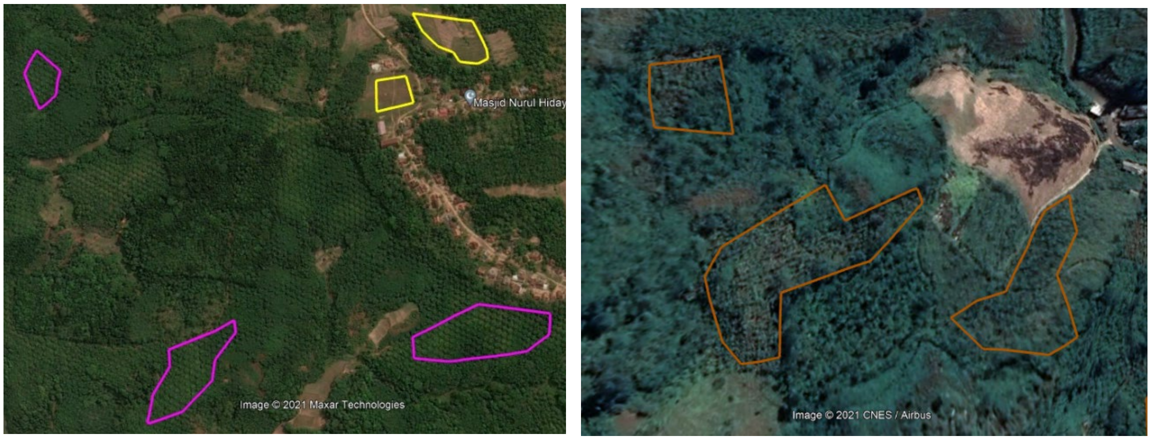



*Figure 25 Polygons of coffee (brown) and no-coffee like pastures and crops (yellow), purple (oil palm)*

From within these polygons, we created 6258 random training samples for the coffee class, and 9229 for the non-coffee land uses (~1:1.5 samples) (Figure 7).

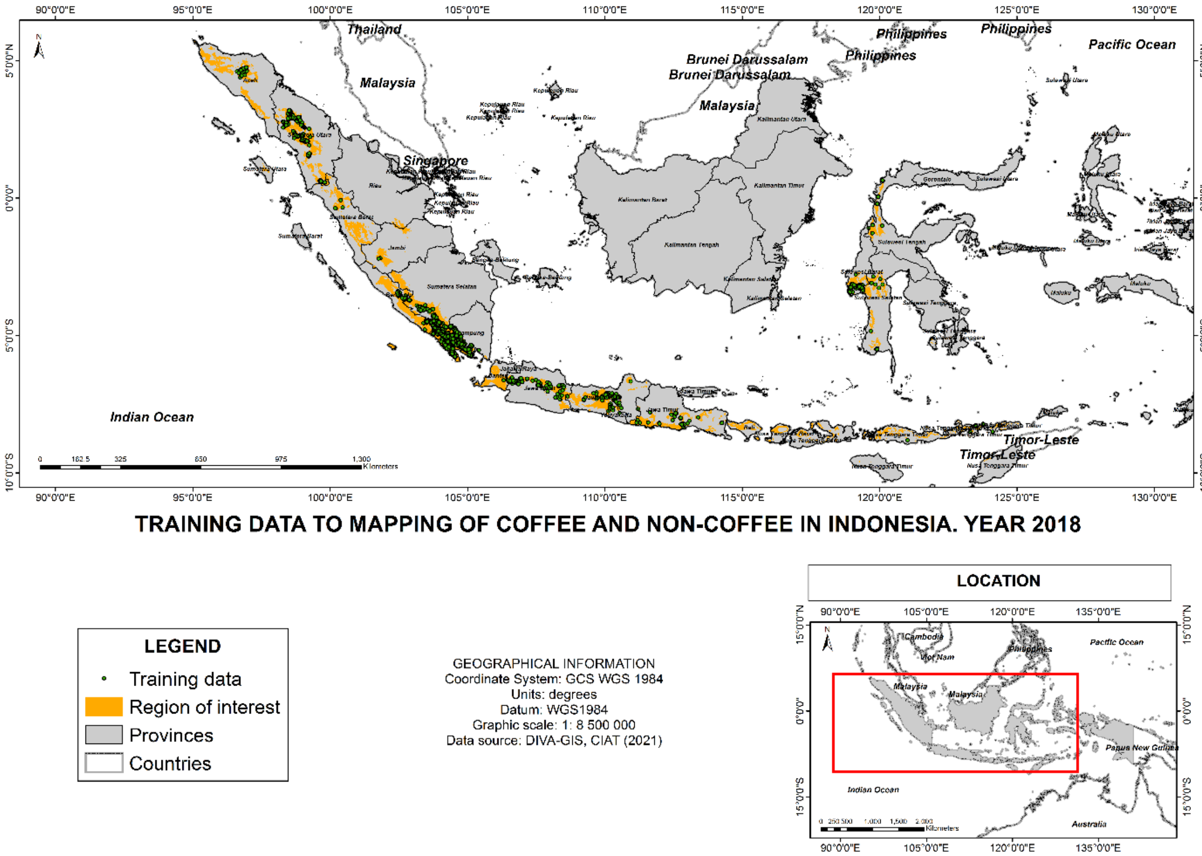



*Figure 26 Training data and region of interest*

## 2.2 Pre-processing of satellite imagery

All optical satellite imagery was pre-processed by applying the region of interest mask, and to reduce the information in the raw dataset of 5000+ available images for 2018.

### *2.2.1 Cloud cleaning*

The satellite images S2-2A were corrected using the product "Cloud Probability" to detect the probability of a pixel being a cloud. The values of this product range from 0 to 100%, where 0 represents a pixel of high quality and 100 corresponds to pixels with clouds. For the region of interest, we selected the pixels with a maximum cloud probability of 20% to include only images without atmospheric errors. Next, we selected the bands B2, B3, B4, B8, B11 and B12 which correspond to Blue, Green, Red, Near Infrared, Short Wave Infrared 1 and Short Wave Infrared 2. To these bands we applied the following steps:

- Cloud probability masking – this filter was applied to every image of the collection, masking and removing all pixels with a cloud probability above 20%.
- Cloud edge masking – in addition to cloud probability masking we removed the cloud edges, which weren't removed before, because for the 10m bands, these often aren't recognized. For this reason, the bands Edge red 4 and water vapor were used.

#### *2.2.2 Vegetation index NDVI*

We calculated the normalized difference vegetation index (NDVI) to reinforce the classification and achieve a differentiation between coffee and other land uses. We scaled each pixel by multiplying with 0.0001 to get an adequate range of values. Next, we used equation 1 to calculate NDVI (Schultz et al. 2016).

$$NDVI = \frac{Near\ infrared - Red}{Near\ infrared + Red} \quad \text{Eq. 1}$$

where Near Infrared corresponds to band B8 and Red to band B4.

### *2.2.3 Sentinel-1 image correction*

The images available in GEE contain the data S1 SAR pre-processed with the toolbox for Sentinel-1 to generate a calibrated and ortho-corrected product. We followed steps to derive the coefficient of retro dispersion for each pixel, starting with the application of the orbita file, followed by the edge noise elimination GRD, thermal noise reduction, and radiometric calibration, and finally terrain correction using el 30m DEM SRTM or DEM ASTER (compare https://bit.ly/385Ujud).

### *2.2.4 GLCM texture indices*

We calculate the co-occurrence matrix of grey levels using the tool glcmTexture in GEE which derives 14 metrics proposed by (Haralick, Shanmugam, and Dinstein 1973) and 4 additional metrics by (Conners, Trivedi, and Harlow 1984). Of the 18 bands produced by the tool, we selected entropy and dissimilarity for the polarizations VH and VV because entropy is part of the features that make the classification algorithm more efficient (Tridawati et al. 2020). The dissimilarity is used to differentiate areas where the pixels regions are not homogeneous in gray levels (Ferreira et al. 2019). The 4 bands used in the classifications model, were VH_dissimilarity, VV_dissimilarity, VH_entropy, and VV_entropy.

### *2.2.5 Reduction of imagery*

After area exclusion and quality filter, we created a consolidated image from all available imagery for the year 2018, consisting of the spectral bands blue, green, red, swir1 and awir2. This 2018 image consists of statistical indicators for each band, including the minimum, maximum, mean, median, and standard deviation. Additionally, the digital elevation data was added, so that the final image consisted of 41 spectral bands.

*Table 6 Spectral bands in the final satellite image consisting of statistical indicators*

| **Name of spectral band** | **Number of bands** | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Digital level** | **Mean** | **Median** | **Standard deviation** | **Minimum** | **Maximum** | **Total** |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| ***Blue*** | 0 | 1 | 1 | 1 | 1 | 1 | 5 |
| ***Green*** | 0 | 1 | 1 | 1 | 1 | 1 | 5 |
| ***Red*** | 0 | 1 | 1 | 1 | 1 | 1 | 5 |
| ***Nir*** | 0 | 1 | 1 | 1 | 1 | 1 | 5 |
| ***Swir1*** | 0 | 1 | 1 | 1 | 1 | 1 | 5 |
| ***Swir2*** | 0 | 1 | 1 | 1 | 1 | 1 | 5 |
| ***NDVI*** | 0 | 1 | 1 | 1 | 1 | 1 | 5 |
| ***Slope*** | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| ***Elevation*** | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| ***VH_entropy*** | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| ***VV_entropy*** | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| ***VH_dissimilarity*** | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| ***VH_dissimilarity*** | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| **Total bands** | | | | | | | 41 |

## 2.3 Classification of land cover and model validation

The land use classification was implemented in the platform RStudio, using the library RGEE to communicate with Earth Engine. This approach enabled the optimization of RandomForest parameters with the R libraries mlr3, RandomForest and RFE. The process consists of the following steps:

1. Read files in GEE: all spatial data is stored in the asset of GEE, such as the pre-processed Sentinel-2A and Sentinel-1 data for 2018 from the previous step, and the training data.
2. Definition of objective: For all training data the classification objective is specified by adding the respective value to each reference location. The values were either coffee, or one of the non- coffee classes.
3. Separation of training data: The data set is split 70/30% into a training and validation dataset.
4. Extraction of data: For each training and testing location the data for each spectral bands is extracted from the imagery.
5. Variable selection introducing variables which are correlated can result in overfitting of the classifier. We therefore used the libraries BORUTA and RFE to select variables. These tools select the variables which most contribute to the classification model, based on the variables importance attribute in Random Forest. Both algorithms suggested to use all 41 variables (Annex I).
6. Cross-validation and tuning: The training data is divided into 5 groups for cross-validation. Different combinations of key tuning parameters for Random Forests (Node size, variables picked (mtry), and number of trees - Table 2) are used and the accuracy on each data subset is evaluated (mlr library):

*Table 7 Parameter spaces tested for Random Forest tuning*

| Number of trees | | mTRY | | Node number | |
|---|---|---|---|---|---|
| Minimum | Maximum | Minimum | Maximum | Minimum | Maximum |
| 500 | 1000 | 5 | 41 | 5 | 41 |

The parameter combination which performed highest according to the accuracy metric was used to train the classification algorithm:

- Mtry = 15
- Node = 5

- Ntree = 1000

7. RandomForest classification: Finally, we used the Random Forest classifier trained in this process to classify the pre-processed summary image in Google Earth Engine. The classifier was tested on the internal cross validation data (Table 3).

*Table 8 Cross validation classification result*

| | Non-coffee | Coffee | class.error (%) |
|---|---|---|---|
| **Non-coffee** | 1784 | 78 | 4.18 |
| **Coffee** | 95 | 1125 | 7.78 |

## 2.4 Post-processing of the classification

Pixel based algorithms, as used here, often produce a “salt and pepper effect”, because each pixel is evaluated in isolation. We applied post procession to the results to reduce the impact of this effect and better generalize the map. Four spatial processing steps were used: majority filter, region groups, cell elimination, and gap filling.

We first applied the function ‘majority filter’. The function replaces cells in a raster based on the majority of their contiguous neighboring cells. The neighborhood consists of a window of 9 pixels and the central pixel will change its value if the majority of its neighbors have the same value. This is illustrated in Figure 8 (Esri, 2016).

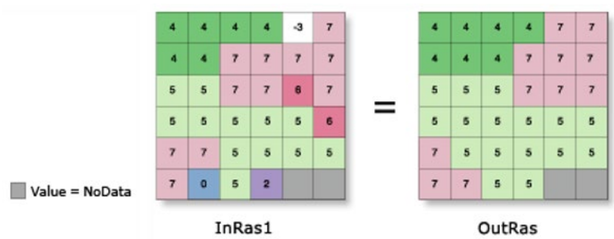



*Figure 27 Majority filter tool illustration (ESRI 2016).*

Next, the region groups filter assigns a unique number to each region. Regions are a contiguous set of cells of the same zone type. When the regions need to be processed separately, each must be identified as a separate entity. The Region Group tool assigns a new value to each region in a raster as shown in Figure 9 (Esri, 2016)

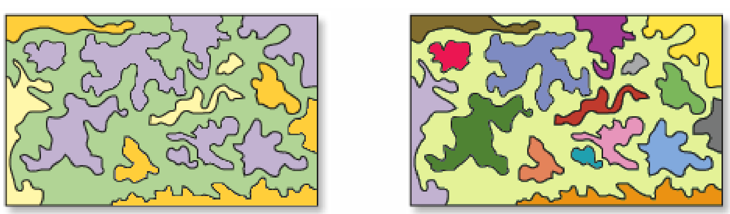


*Figure 28 Region groups. Left: regions separated by a single value, Right: Disconnected regions with a unique value (ESRI 2016).*

To reduce spurious grid cells, we assigned NoData values at region groups equivalent to <0.50ha. The tool SetNull returns NoData if this conditional evaluation is true, and returns the value specified by the original raster if it is false (Figure 10).

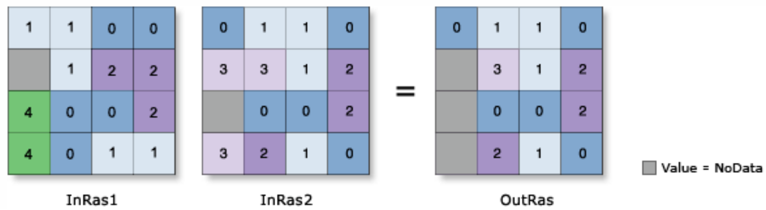



*Figure 29 Elimination of spurious groups using SetNull in Arcgis (ESRI 2016)*

Finally, the information for these grid cell groups was filled with the values from neighboring fields with the tool Nibble from ArcGis. Nibble replaces cells of a raster corresponding to a mask with the values of the nearest neighbors (Figure 11).

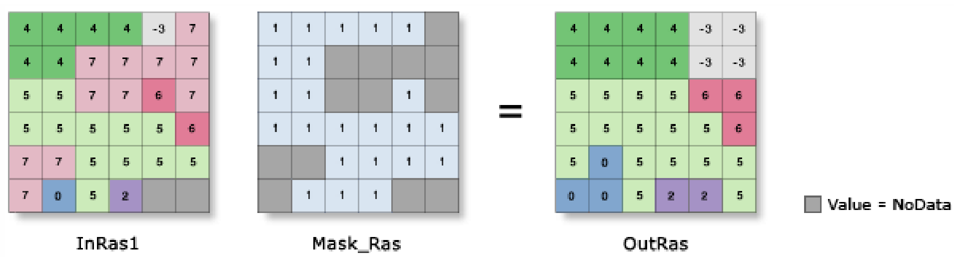



*Figure 30 Gap filling using the tool Nibble in ArcGis (ESRI 2016).*

## 2.5 Validation

We validated the result map by (1) comparing reported area in national statistics by municipality and modelled area, (2) by creating a reference pixel database on high resolution satellite data, and (3) in stakeholder workshops[1]. We identified regions of low classification performance and repeated the steps in "3.1.3 - Training data" and the following once.

### *2.5.1 Comparison of reported and mapped area*

We compared the modeled area according to the classification into coffee/non-coffee, with reported planted area statistics of coffee for the year 2018 to identify regions with low model accuracy. Area statistics were reported at municipal level, so that we estimated modeled area for each municipality. We then calculated the difference between the two data sources. We differentiated municipalities with underestimation of area (< 100% of reported area), low error (100 – 300% of reported area), moderate error (300 – 500%), high error (500 – 700%) and very high error (>700% of area).

The resulting map was evaluated to prioritize regions for additional training data sampling (Figure 12). We further summarized it by counting the number of municipalities with very low, low, moderate, high, and very high error rates (Table 4). Following the visual inspection of the map, and review of the table, we prioritized Jawa province for additional sampling. With the added training data, we repeated the classification steps. After the first repetition, the map of coffee/non-coffee had 30 municipalities less in the "high and very high error" categories (Figure 12 and Table 4).

*Table 9 Number of municipalities in each classification error group between initial and final model*

| | Error rate | | | | |
|---|---|---|---|---|---|
| | Very low | Low | Moderate | High | Very high |
| **Initial model** | 34 | 44 | 28 | 7 | 50 |

[1] We included the reference to a validation workshop to illustrate the purpose. The workshop still needs to be held.

| | | | | | |
|---|---|---|---|---|---|
| **Final model** | 105 | 20 | 10 | 11 | 16 |

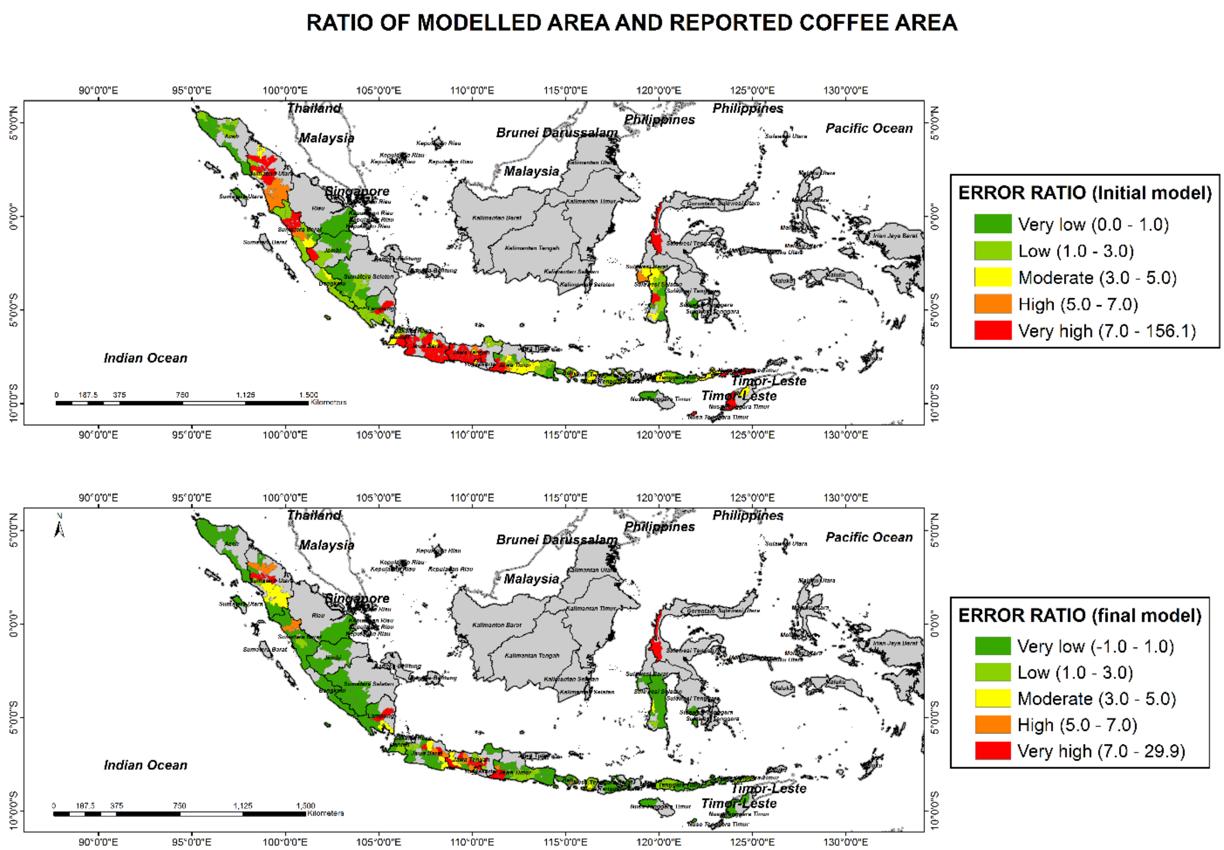



*Figure 31 Ratio of modelled area and reported coffee area.*

Visual inspection of the final map showed that distribution of error no longer showed a clustered spatial distribution. Most municipalities now showed an acceptable error (<300% of reported area).

### 2.5.2 Reference data from high-resolution satellite data

We used stratified sampling of reference data from high resolution satellite imagery to validate classification accuracy following methods suggested by (Olofsson et al. 2014). We used the SEPAL tool (https://sepal.io) to generate a random stratified location sample for validation. The tool accounts for the expected prevalence of land use classes, and the classification result of the preliminary map. The tool creates samples from the entire region of interest for each land use class. In addition to prevalence, sampling depends on the expected accuracy and desired confidence intervals. In our case, the coffee class is considered to be a relatively rare class, compared to other land uses in the region of interest. For this reason, we defined the expected accuracy to be 60% and for non-coffee land use 90%. We validated 318 locations, of which 108 were for coffee, and 210 for non-coffee. The spatial distribution can be seen in the following map (Figure 13)[2].

[2] Within the provided time frame, we were unable to evaluate a greater validation data set. The correct sample number should be 562 non-coffee samples and 315 coffee locations.

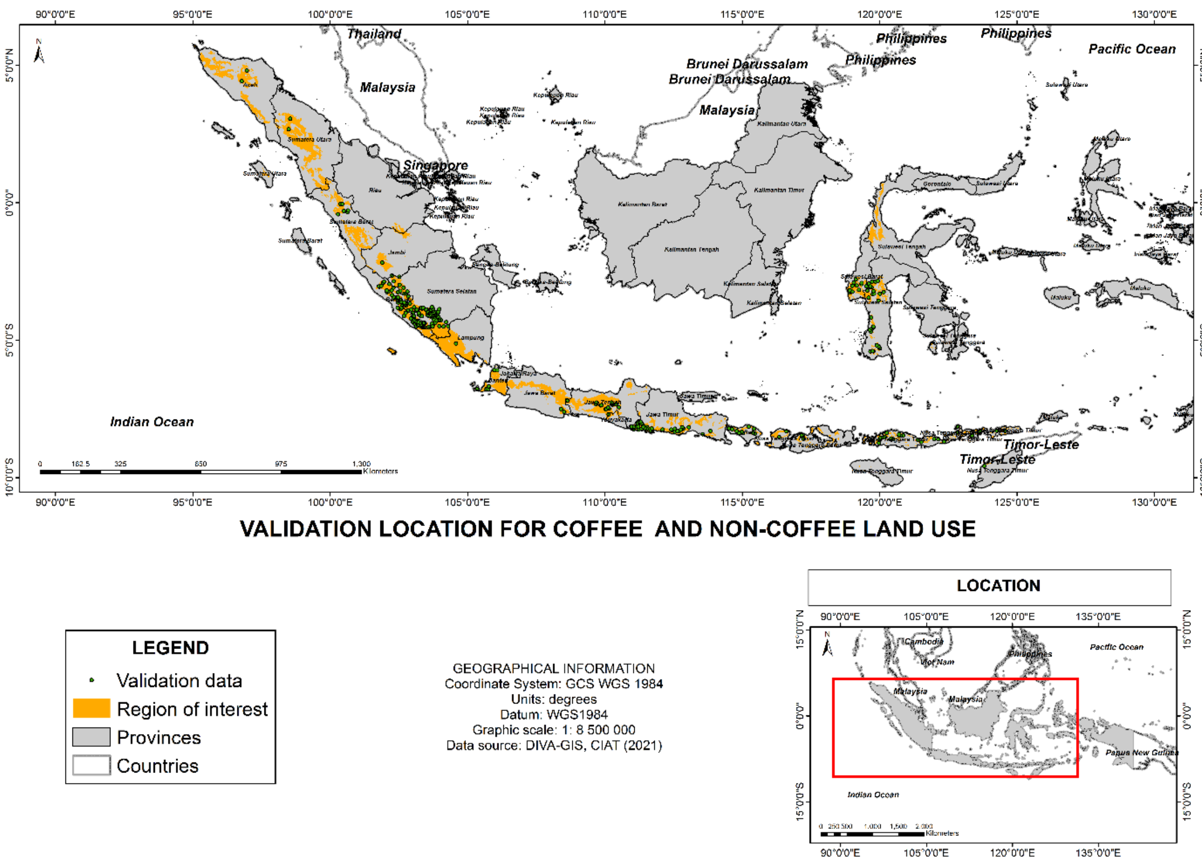



*Figure 32 Validation locations for coffee and non-coffee land use by stratified random sampling.*

For 308 references locations satellite imagery of sufficient quality was available to evaluate with confidence whether a location was coffee or non-coffee. The coffee region in Indonesia does not have a good coverage with images for years close to 2018 in the region of interest, reducing the confidence in the correct identification of coffee areas in the region of interest.

Using these data, we created a confusion matrix after generalization of at least 0.5ha.

# 3 Results

The first step of the map quality assessment was to generate a confusion matrix. From this matrix, we extracted a set of indicators including, the producer and user accuracy for coffee and non-coffee classes as well as the overall accuracy (Annex 2).

The results show that the producer accuracy is 82.1% for the coffee class, which is equivalent to omission errors of 17.9%. On the other hand, the non-coffee producer accuracy was significantly higher, reaching 85.4%. Coffee being a relatively rare class in comparison with the non-coffee one, it was expected to observe a greater probability of a false negative.

The user accuracy represents the probability that an area predicted to be of a certain class really is that class. Our result show that the highest user accuracy was obtained after generalization of 0.5ha with a value of 72.2% and 92.1 for coffee and non-coffee respectively. Indonesia

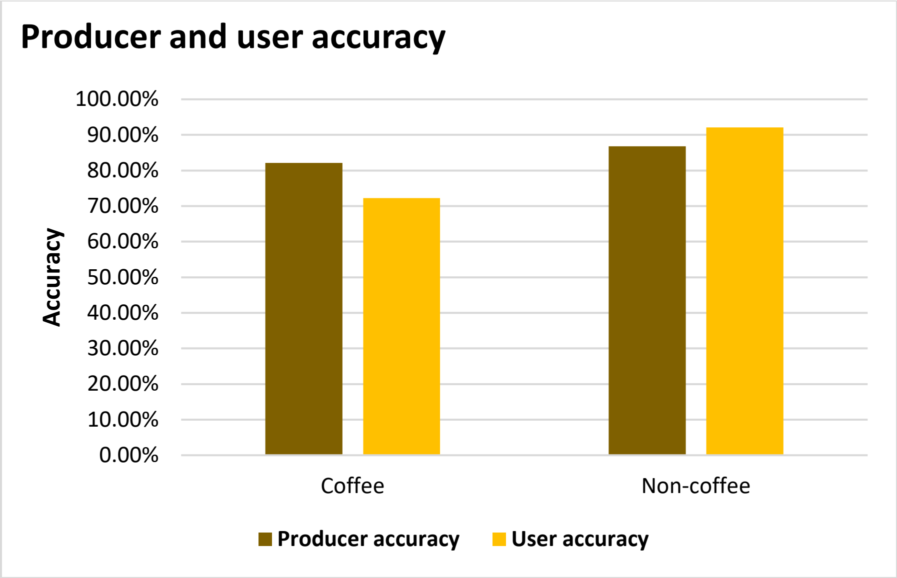



*Figure 33 Producer and user accuracy for different levels of generalization*

The provinces identified with the largest coverage of coffee plantation are Sumatera Utara, Sumatera Selatan, Jawa Tengah, Lampung, and Jawa Timur, with 24.2%, 13.8%, 11.4%, 10.7, and 9.5% of their land used for coffee production respectively. On the other hand, the provinces with the lowest percentages are Riau (0.00%), Sulawesi Tengara (0.00%), Jambi(0.28%), Yogyakarta (0.36%), and Sulawesi Tengah (0.42%).

Based on the DGEC data, we produced the error rate by province (Figure 16). In most provinces our results are estimate more coffee area than officially reported. On the other hand, especially in Aceh and Jambi, we estimated relatively less.

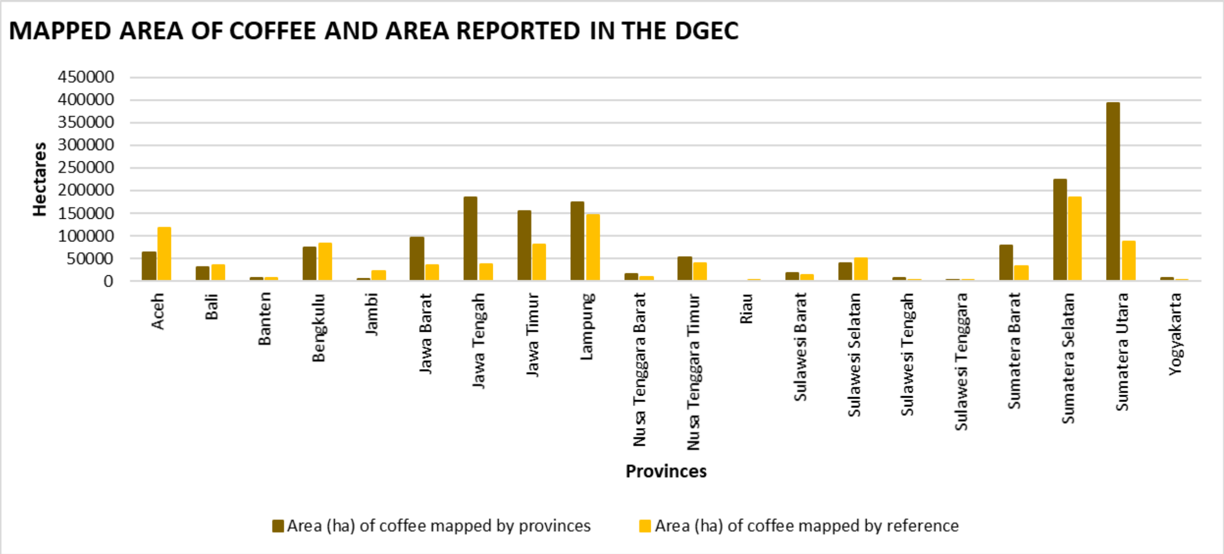



*Figure 34 Reported (EVA - MARD) and modelled area by department.*

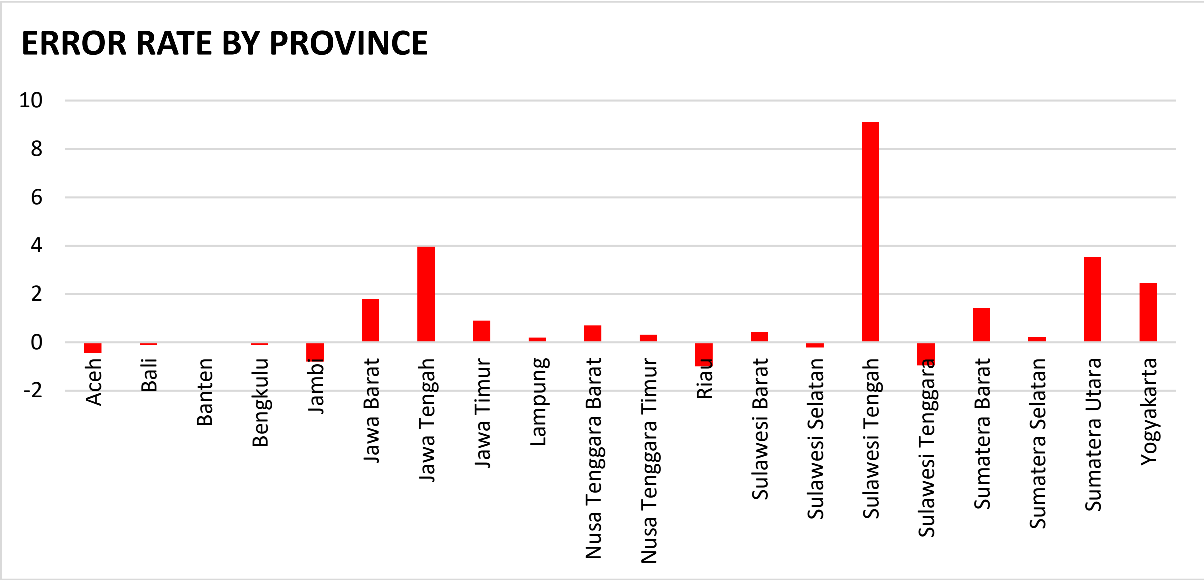



*Figure 35 Error rate by department*

Finally, Figure 17 shows the spatial distribution of coffee and non-coffee area in Indonesia for 2018, with an overall accuracy of 85.4%.

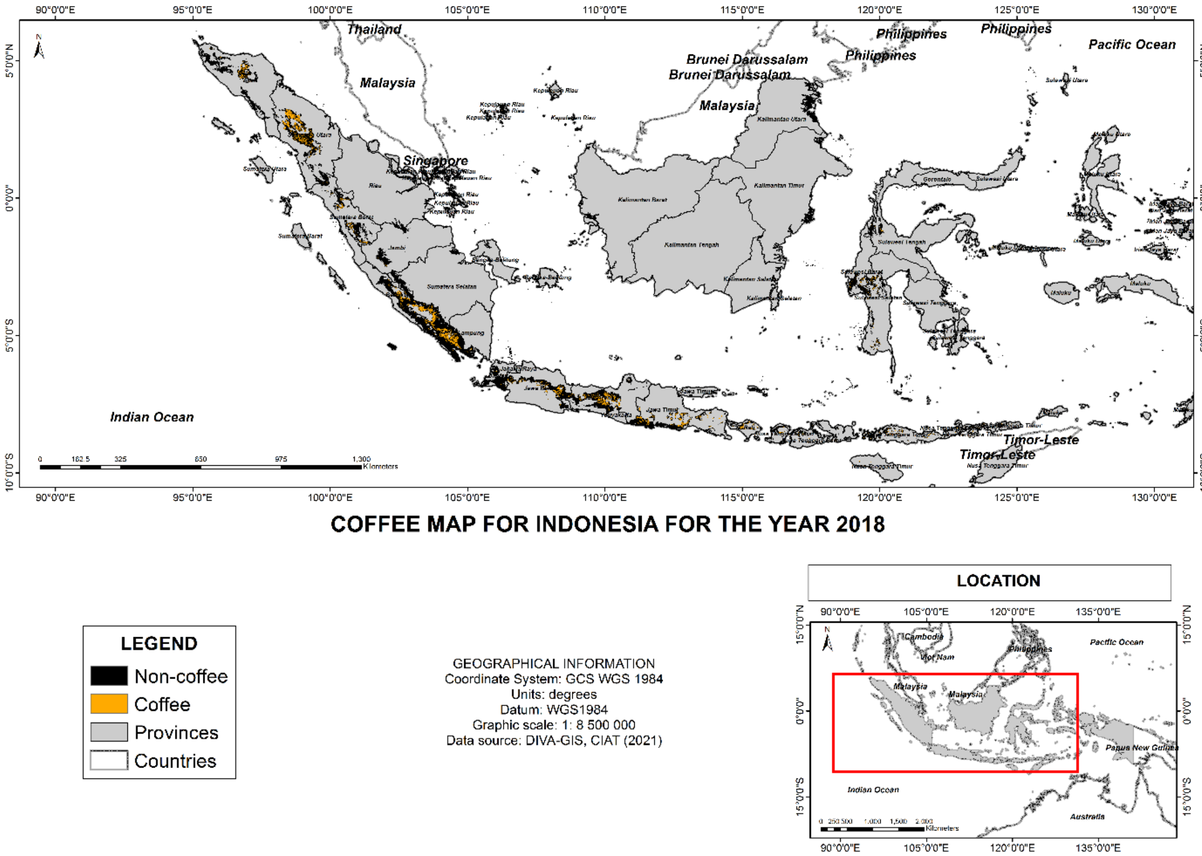



*Figure 36 Final coffee map for Indonesia for the year 2018.*

## 4 Discussion

Knowledge about the distribution of coffee production is a valuable resource to improve data driven decisions support. We demonstrated how precise maps of coffee production can be produced at country scale from freely available remote sensed data. The resulting overall accuracy of 85.4%, and the user accuracies of 72.2% on the coffee class, and 92.1% on the non-coffee class, are in line with previously reported accuracies (Hunt et al. 2020). However, previously, mapping was either restricted to areas where coffee production can be expected to be a dominant land use class.

The reported accuracy could be subject to change with two changes, potentially implemented at a later stage: addition of more validation data, separation of validation data by polygons. Within the available time we were unable to process the required number of validation sample following (Olofsson et al. 2014). Furthermore, model tuning and variable prioritization used a cross-referencing process which drew from within the same polygons. Spatial autocorrelation may inflate statistics in such circumstances. A separation into training and testing data by polygon may change cross-validation statistics.

A key innovation of our work is the use of climate suitability data to delineate the region of interest, making a whole-country mapping computationally feasible. Recent country scale mapping for coffee in Vietnam (Maskell et al. 2021) and cocoa in Ghana and Ivory coast (Abu et al. 2021) showed lower user accuracies, despite using similar data and methods.

Coffee being a relatively rare class in comparison with the non-coffee one, it was expected to observe a greater probability of a false positive (non-coffee area being identified as coffee). This is leading to an overestimation of coffee areas in most provinces, when compared to the data from DGEC. The results however stay within acceptable limits with a producer accuracy of 80% for the coffee class and a user accuracy above 90% for the non-coffee class.

A key challenge was the generation of sufficient training data. Available data sources were not collected for the purpose of informing remote sensing. Geo-references of coffee commonly describe supply chains and may include farm houses, primary processing facilities or collection stations. Such data is valuable to describe flows of coffee and to identify broad origins, but cannot be used to train remote sensing classification models. Even plot-level data, collected to estimate coffee area, remotely observe pest and disease outbreaks or similar uses was not fit for this purpose as adjacent forest may be included, riparian areas or minor patches of other land cover. We therefore had to invest substantial effort into generating clean data from such initial reference locations.

Despite the expert generation of polygons, both training and validation data may contain erroneous land use classes. Coffee production in Indonesia is highly dispersed on small plots, and often intercropping is practiced. This makes them hard to differentiate from other land uses. A likely explanation of the relative overestimation of coffee area is the difficulty to differentiate coffee from other agroforestry crops or adjacent degraded forest.

In this work, we developed a novel approach to reduce the scope of the mapping area based on coffee suitability maps and the inclusion of environmental variables in the analysis. This approach has been shown to be very effective to serve as a first filter to pinpoint the areas the need a precise mapping and discard large areas where coffee is very unlikely to be cultivated, thus saving a large amount of processing time.

Our mapping is not the first attempt at whole country mapping of coffee areas, but likely the most accurate approach available to date. Previous work was largely reliant on subnational production statistics, sometimes in combination with environmental or satellite data (Monfreda, Ramankutty, and Foley 2008; You et al. 2014). Such approaches were sufficiently accurate to inform global analysis, but not accurate enough for decision making by stakeholders (Bebber, Castillo, and Gurr 2016; Eriyagama, Chemin, and Alankara 2014).

Our data is similar in precision as other remote sensed data, such as weather observation, deforestation monitoring or general land use maps. Our mapping can therefore be useful to develop added value services. Coffee production can be beneficial for biodiversity and carbon stocks, if it spares primary forest, and is established in previously degraded landscapes (Martin et al. 2020).The combination of our coffee map and deforestation monitoring will make zero-deforestation commitments by the sector more feasible to implement. The combination with other land cover data can inform the identification of areas of potential coffee expansion. In addition, climate services for coffee will benefit from improved risk and productivity assessments.

To improve the results, approaches, based on convolutional neural networks, could be applied to better take into account the structure of the canopy. The pixel-based analysis used in this study, Random Forest, does not use the full potential of high-resolution imagery such as Sentinel 1 and Sentinel 2. Future work should investigate more this type of deep-learning approach to refine the current map.

## Section 2: HOTSPOT ANALYSIS

- Methodology
- Colombia ReadMe
- Indonesia ReadMe

# HOTSPOT ANALYSIS: METHODOLOGY

**Comparison of Administrative Units' Contribution of Commodity-associated Tree Cover Loss to Priority Forested Ecoregions**

## Overview

This analysis is intended to model the relationship between observed tree cover loss (i.e., deforestation) and coffee production in the pilot countries Colombia and Indonesia, ultimately prioritizing jurisdictions for engagement/mitigation efforts. Due to differences in the schema of data available in both countries, the approach to modeling this relationship takes different approaches in the two pilot countries. Both models are based on a method of jurisdictional ranking previously developed by the authors that prioritizes jurisdictions based on a combination of the quantity of coffee-associated deforestation and local rates of endemic species richness that is potentially impacted by the deforestation (Follett and Slay 2020). For the purposes of this project tree cover loss and deforestation are synonymous and defined as conversion from forest to coffee.

## Methods and Data

### Background

This analysis is an evolution of a previous analysis conducted by WRI on behalf of the Tropical Forest Alliance (TFA), whose purpose was to quantify the contribution of several top commodities to current rates of tropical deforestation (Goldman et al. 2020). The model used the MapSPAM model's data on physical area under cultivation to determine what portion of the deforestation in each location should be attributed to each commodity of interest, and perhaps more importantly, what portion should not be attributed to those commodities (Wood-Sichra et al. 2016). MapSPAM spatially disaggregates statistics on physical area under cultivation in a certain administrative area. These statistics are modeled as a global grid of approx. 10km cells, each with a quantity of its footprint that is used in the production of each of the 42 crops/crop categories that are modeled by MapSPAM data.

For the TFA report, WRI used these numbers to derive the portion of that cell's physical area devoted to production of a given commodity and used that portion of the sum of deforestation in the cell to attribute to each of the MapSPAM crops. To justify attributing all of a cell's deforestation to commodity expansion, the model was constrained to MapSPAM cells where the dominant driver for deforestation classified by Curtis et al. (2018) was either Commodity-Driven or Smallholder/Shifting Agriculture, the two classes of drivers in this model associated with agricultural production.

Expanding on the work of Goldman et al., the authors previously created a model to "quantify the value of intervention in [a] jurisdiction in relation to how much rare biodiversity is at risk from the jurisdiction's deforestation" (Follett and Slay 2020). This approach used the World Wildlife Fund's *Global 200 Priority Ecoregions* dataset (Olsen & Dinerstein, 2002), a precursor to Dinerstein et al. 2016, to constrain the analysis to ecoregions whose conservation status was defined as 'critical or endangered' or 'vulnerable', excluding the only other class of conservation status: 'relatively stable/intact'. Within these priority forest ecoregions, the model computed the portion of the ecoregion's total tree cover removed in a given jurisdiction. These portions are summed across all ecoregions that intersect a jurisdiction that is being scored. As a result, the model favors intervention in areas where a jurisdiction is contributing significant amounts of deforestation to multiple priority ecoregions, meaning that engagement there has the potential to protect and conserve multiple priority ecoregions.

### Model Methodology

For the purposes of this model, deforestation was associated specifically with coffee production. During this project a high-resolution map of coffee production in Colombia and Indonesia was produced by CIAT (Castro et al. 2021), supplanting the use of the coarse crop distribution information provided by

MapSPAM. With this higher resolution coffee production data, a more straightforward approach to associating deforestation to commodity production is possible by taking the intersection of the coffee map and the tree cover loss map. To conduct the analysis at this higher resolution, the Hansen et al. loss data is resampled to this 10m resolution so that direct comparisons can be made between the two. As both the coffee distribution map and the tree cover loss data are a binary classification of coffee/not coffee, tree cover loss/not tree cover loss, we select all the pixels that are both coffee and tree cover loss and use that as our map of coffee associated deforestation.

For the purposes of this analysis, Hansen et al. tree cover loss data was filtered to the 5 most recent years of loss (2016-2020 inclusive). The intent of using this subset of the loss data is to model the spatial distribution of future loss more accurately by constraining the input loss to recent years. As loss rates exhibit some degree of temporal autocorrelation, the assumption is that the most recent 5 years will more closely resemble the spatial distribution of future loss than the entirety of the available loss data which goes back to 2000. We did not include updated loss drivers in the model (unlike WRI's TFA report) as the assumption was made that coffee production was the driver of forest loss in areas of co-occurrence.

**Jurisdictional Prioritization – Indonesia**

Jurisdictional prioritization of efforts aimed at mitigating or addressing coffee-associated deforestation have the potential to increase the return on investment by targeting areas with disproportionate impacts on natural landcover and local biodiversity. Though the relationship between coffee and tree cover loss in this analysis is conducted at a higher resolution than the previous TFA report (Goldman et al., 2020), the datasets for quantifying tree cover and delineating priority ecoregions are the same as (Follett & Slay, 2020). This approach uses a similar scoring algorithm to that of Follett and Slay 2020, where a jurisdiction's score is a summation of its contribution to priority forested ecoregions intersecting it. Given functions D(x) to represent the area (expressed in hectares) of coffee-associated deforestation in region x and TC(x) to represent the hectares of tree cover in x, the scoring algorithm developed by Follett and Slay can be expressed as

$$Score(A_i)=\sum_{j=1}^{n}\frac{D(A_i\cap E_j)}{TC(E_j)}$$

where $A_i$ represents the $i^{th}$ element in set A of all administrative units to be scored, $E_j$ represents the $j^{th}$ ecoregion in set E of all priority ecoregions that intersect $A_i$. Therefore, the prioritization score of a given jurisdiction $A_i$ is the sum of the ratios of total area of commodity associated deforestation in the intersection of $A_i$ and $E_j$ to the total area of tree cover in the entirety of ecoregion $E_j$ circa 2010 for all ecoregions intersecting $A_i$. The intent of this ratio is to quantify a jurisdiction's contribution to the deforestation of its constituent priority ecoregions.

**Jurisdictional Prioritization – Colombia**

The scoring approach for Colombia differed significantly from the original version of this analysis as there was both highly detailed landcover/ecosystem data available for the country, as well as the high-resolution coffee distribution map generated by CIAT. The high-resolution coffee map takes the place of the MapSPAM model for commodity distribution in the original approach by Follett and Slay 2020 and the Colombian ecosystems dataset takes the place of the priority ecoregions and landcover data. Because of differences between these datasets and the data used for the original approach, the question being asked was adapted slightly. While the Dinerstein 2017 dataset included data on the threatened status of an ecoregion, the Ideam 2017 dataset does not contain that information (Dinerstein et al., 2017, Ideam et al., 2017). It does, however, provide landcover classification and endemic species richness data, along with characterizing all landcover as natural or 'transformed' (i.e., non-natural), all at a scale of 1:100,000. The natural/'transformed' classification is an incredibly valuable attribute as it allows the focus of the model to be tuned specifically to coffee's contribution to land cover conversion rather than the map of overall tree cover used in the Indonesian analysis and prior versions of this model.

This model looks at the portion of a biotic unit's non-natural landcover that is coffee, and weights that by the number of unique endemic species in the intersection of a given jurisdiction and ecoregion. This is aggregated to the municipal level via summation of this ratio in all the constituent ecoregions within each jurisdiction, like the approach in the Indonesian model above.  Thus, the score of an administrative unit $A_i$ can be expressed as

$$score(A_i)=\sum_{j=1}^{n}\frac{D(A_i\cap E_j)}{T(A_i\cap E_j)}\cdot S(A_i\cap E_j)$$

where D(x) represents the area of coffee-associated deforestation in area x, T(x) represents the total area of transformed (non-natural) land cover in x and S(x) represents the number of unique endemic species found in x.

**Limitations**

These models are purely focused on the geophysical variables of coffee production, endemic species richness and landcover types.  This analysis does not consider political or economic factors when prioritizing jurisdictions for engagement. The data for deforestation was restricted to the 5 most recent years of loss data: 2016-2020 (inclusive).

These models do not establish any causal relationship between coffee production and deforestation.  This relationship is purely based on co-presence, i.e., both factors being detected by remote sensing methods in the same location.

**Citations**


Castro, F. and Bunn, C. (2021). Coffee suitability mapping for coffee in Colombia and Indonesia under past and future conditions. Alliance of Biodiversity and CIAT. Cali, Colombia.

Curtis, P.G., Slay, C.M., Harris, N.L., Tyukavina, A., and Hansen, M. (2018). Classifying drivers of global forest loss. Science. 361. 1108-1111. https://doi.org/10.1126/science.aau3445

Dinerstein, E., Olson, D., Joshi, A., Vynne, C., Burgess, N.D., Wikramanayake, E., Hahn, N., Palminteri, S., Hedao, P., Noss, R., Hansen, M., Locke, H., Ellis, E.C., Jones, B., Barber, C.V., Hayes, R., Kormos, C., Martin, V., Crist, E., Sechrest, W., Price, L., Baillie, J.E.M., Weeden, D., Suckling, K., Davis, C., Sizer, N., Moore, R., Thau, D., Birch, T., Potapov, P., Turubanova, S., Tyukavina, A., de Souza, N., Pintea, L., Brito, J.C., Llewellyn, O.A., Miller, A.G., Patzelt, A., Ghazanfar, S.A., Timberlake, J., Klöser, H., Shennan-Farpón, Y., Kindt, R., Lillesø, J.B., van Breugel, P., Graudal, L., Voge, M., Al-Shammari, K.F., Saleem, M. (2017). An Ecoregion-Based Approach to Protecting Half the Terrestrial Realm. BioScience Volume 67, Issue 6. Pages 534–545. https://doi.org/10.1093/biosci/bix014

Follett, F., Slay, C.M. (2020). Ranking Priority Jurisdictions with Recent Deforestation for Specific Commodities. The Sustainability Consortium. https://sustainabilityconsortium.org/download/follett-slay-2022-comparison/

Goldman, E., Weisse, M.J., Harris, N., and Schneider, M. (2020). Estimating the Role of Seven Commodities in Agriculture-Linked Deforestation: Oil Palm, Soy, Cattle, Wood Fiber, Cocoa, Coffee, and Rubber. Technical Note. World Resources Institute. Washington, DC. https://doi.org/10.46830/writn.na.00001

Instituto de Hidrología, Meteorología y Estudios Ambientales (Ideam), Instituto de Investigación de Recursos Biológicos Alexander von Humboldt (Instituto Humboldt), Instituto de Investigaciones Marinas y Costeras José Benito Vives de Andréis (Invemar) e Instituto Geográfico Agustín Codazzi (IGAC). Memoria técnica. (2017). Mapa de ecosistemas continentales, costeros y marinos de Colombia (MEC), escala 1:100.000. 170 pp.

# HOTSPOT ANALYSIS: COLOMBIA READ ME

Colombia Coffee Reforestation Prioritization Model

## About

**Author:** Forrest Follett (forrest.follett@sustainabilityconsortium.org)
**Software Used:** ArcGIS Pro 2.9.2
**Model Intent:** This model ranks jurisdictions based on the potential return on investment of reforestation projects specific to the coffee sector. This model uses the unidades bioticas (biotic units) from the Colombian MEC dataset (listed below) as the ecological unit of analysis, intersected with the second level administrative units (municipios) to create the base unit of analysis.
**Output File:** /The Sustainability Consortium/GIS/Prioritized_municipios_v2_5_18.zip (GIS)
**Output Fields:**

- refstat_sc: Score assigned to each jurisdiction by the prioritization model (scaled from 0-100), used to rank the jurisdictions. This score is a unitless distribution intended to indicate the relative value of conducting reforestation activities in a jurisdiction's coffee landscapes compared to other coffee-growing jurisdictions.
- ctcl_ha – Area (in hectares) of the overlap of 'recent treecover loss' (2016-2020 inclusive) with CIAT coffee map

## Datasets:

*CIAT coffee map*
Sharepoint path: /Walmart Folder for ALL/Component 2/GIS_Colombia/CIAT_COFFEE_MAP
Description: The coffee suitability dataset produced by CIAT was included in Step 1 for users to be able to select if they wanted to view restoration opportunities exclusively within regions suitable for coffee production. The method to produce this dataset consisted of five general steps: initial training data collection, satellite data pre-processing, algorithm training, data classification and results post processing, and fourth, results validation. Validation demonstrated an 85.4% overall accuracy and a user accuracy of 72.2 and 92.1% for coffee and non-coffee land use respectively. For additional information on detailed steps of the method, visit the Report: CIAT Coffee Map in 2018 for Colombia Technical Report_EN.docx.
Citation: Castro, Fabio and Bunn, Christian; Coffee suitability mapping for coffee in Colombia and Indonesia under past and future conditions; 2021; Alliance of Biodiversity and CIAT; Cali, Colombia.

*Global 2010 Tree Cover (30m)*
Download link: https://glad.umd.edu/dataset/global-2010-tree-cover-30-m
Description: Global tree cover data (treecover2010) are per pixel estimates of circa 2010 percent maximum (peak of growing season) tree canopy cover derived from cloud-free annual growing season composite Landsat 7 ETM+ data. A regression tree model estimating per pixel percent tree canopy cover was applied to annual composites from 2000 to 2012 inclusive (Hansen and others, 2013). Data gaps and noise from individual years were replaced using multi-year median values. First, a median from annual tree canopy cover values from 2009-2011 was used to estimate 2010 tree cover. For pixels still lacking an estimate, the median calculation was expanded to include tree cover values from 2008-2011, then from 2008-2012. Any remaining gaps were filled with tree canopy cover values derived from a regression tree model using all growing season Landsat ETM+ data as inputs. The resulting layer represents estimated maximum tree canopy cover per pixel, 1-100% for the year 2010 in integer values (1-100).
Citation: Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O., and Townshend, J.R.G., 2013, High-Resolution Global Maps of 21st-Century Forest Cover Change: Science, v. 342, no. 6160, p. 850-853. http://www.sciencemag.org/content/342/6160/850.abstract.

*Global Forest Change*
Sharepoint path: Download link in citation
Description: The Global Forest Change Product provides results from time-series analysis of 654,178 Landsat images in characterizing forest extent and change product. For definitions of Forest extent and change refer to Hansen et al., 2013.
Citation: Hansen, M. C., P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend. 2013. "High-Resolution Global Maps of 21st-Century Forest Cover Change." Science 342 (15 November): 850-53. Data available on-line from: http://earthenginepartners.appspot.com/science-2013-global-forest.

*Mapa de ecosistemas continentales, costeros y marinos de Colombia (MEC)*
Sharepoint path: /Walmart Folder for ALL/Component 2/GIS_Colombia/Mapa_Ecosistemas_Colombia_2017
Description: La elaboración del mapa a escala 1:100.000 se hizo con el objetivo de identificar, clasificar y caracterizar los ecosistemas para entender el territorio; por tanto, este proceso trasciende hacia la consolidación y el mantenimiento de un trabajo interinstitucional enfocado en el monitoreo de ecosistemas, en el que se integren características ecológicas, económicas y sociales para orientar las metas de gestión de lo nacional a lo local.
Translated Description: The elaboration of the map at a scale of 1:100,000 was made with the objective of identifying, classifying and characterizing the ecosystems to understand the territory; therefore, this process transcends towards the consolidation and maintenance of inter-institutional work focused on monitoring ecosystems, in which ecological, economic and social characteristics are integrated to guide management goals from the national to the local.
Citation: Instituto de Hidrología, Meteorología y Estudios Ambientales (Ideam), Instituto de Investigación de Recursos Biológicos Alexander von Humboldt (Instituto Humboldt), Instituto de Investigaciones Marinas y Costeras José Benito Vives de Andréis (Invemar) e Instituto Geográfico Agustín Codazzi (IGAC). Memoria técnica (2017). Mapa de ecosistemas continentales, costeros y marinos de Colombia (MEC), escala 1:100.000. 170 pp.

*Mapa entidades territoriales de la República de Colombia*
Download Link: https://www.colombiaenmapas.gov.co
Description: Entidad territorial fundamental de la división político - administrativa del Estado, con autonomía política, fiscal y administrativa dentro de los límites que le señalen la Constitución y las leyes de la República y cuya finalidad es el bienestar general y el mejoramiento de la calidad de vida de la población en su respectivo territorio.
Se representan sobre cartografía del IGAC acorde con la Ley 1447 de 2011 y su Decreto Reglamentario 1170 de 2015. Para los Distritos la definición y modificación de sus límites está estipulado en la Ley 1617 de 2013. Las áreas No Municipalizadas, hacen parte de la división territorial, pero no son entidades territoriales (artículo 285 y 286 de la Constitución Política de Colombia, 1991). La categorización de cada municipio se da de acuerdo a la Ley 617 de 2000.
Translated Description: Fundamental territorial entity of the political-administrative division of the State, with political, fiscal, and administrative autonomy within the limits indicated by the Constitution and the laws of the Republic and whose purpose is the general welfare and the improvement of the quality of life of the population in their respective territory.
They are represented on IGAC cartography in accordance with Law 1447 of 2011 and its Regulatory Decree 1170 of 2015. For Districts, the definition and modification of their limits is stipulated in Law 1617 of 2013. Non-Municipalized areas are part of the division territorial, but they are not territorial entities (articles 285 and 286 of the Political Constitution of Colombia, 1991). The categorization of each municipality is given according to Law 617 of 2000.
Citation: IGAC. 2021. Mapa entidades territoriales de la República de Colombia. https://www.colombiaenmapas.gov.co

# HOTSPOT ANALYSIS: INDONESIA READ ME

Indonesia Coffee Reforestation Prioritization Model

## About:

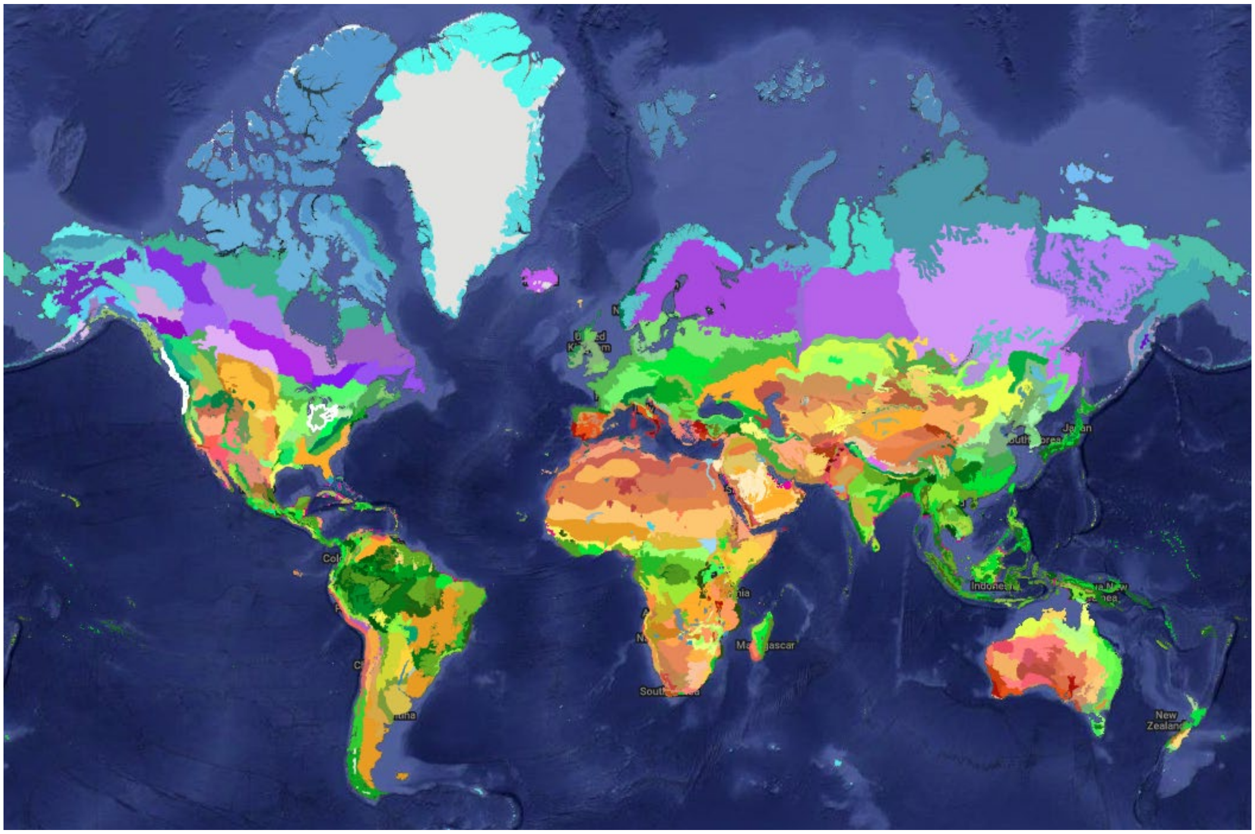
**Author:** Forrest Follett (forrest.follett@sustainabilityconsortium.org)
**Software Used:** ArcGIS Pro 2.9.2
**Model Intent:** This model ranks jurisdictions based on the potential return on investment of reforestation projects specific to the coffee sector. The model seeks to maximize the area of coffee-associated deforestation a jurisdiction contributes to an ecoregion as a fraction of the treecover extent of that ecoregion (c. 2010). Each jurisdiction's score is the sum of that fraction across all the ecoregions found in that jurisdiction. Summing the scores for each ecoregion found in a jurisdiction skews the model to those locations that are deforesting multiple jurisdictions, which facilitate broader impact on a variety of ecological resources.
**Output File:** /The Sustainability Consortium/GIS/prioritized_kabupaten_5_18.zip
**Output Fields:**

- tcl_ha – Area (in hectares) of 'recent treecover loss' (2016-2020 inclusive)
- ctcl_ha – Area (in hectares) of the overlap of 'recent treecover loss' (2016-2020 inclusive) with CIAT coffee map
- **score_scl -** Score assigned to each jurisdiction by the prioritization model (scaled from 0-100), used to rank the jurisdictions. This score is a unitless distribution intended to indicate the relative value of conducting reforestation activities in a jurisdiction's coffee landscapes compared to other coffee-growing jurisdictions.

## Datasets:

*CIAT coffee map*
Sharepoint path: /Walmart Folder for ALL/Component 2/Indonesia_Analysis_GIS_Indonesia/CIAT_IDN_coffee_kk.zip
Description: The coffee suitability dataset produced by CIAT was included in Step 1 for users to be able to select if they wanted to view restoration opportunities exclusively within regions suitable for coffee production. The method to produce this dataset consisted of five general steps: initial training data collection, satellite data pre-processing, algorithm training, data classification and results post processing, and fourth, results validation. Validation demonstrated an 85.4% overall accuracy and a user accuracy of 72.2 and 92.1% for coffee and non-coffee land use respectively. For additional information on detailed steps of the method, visit the Report: CIAT Coffee Map in 2018 for Indonesia Technical Report_EN.docx.
Citation: Castro, Fabio and Bunn, Christian; Coffee suitability mapping for coffee in Colombia and Indonesia under past and future conditions; 2021; Alliance of Biodiversity and CIAT; Cali, Colombia.

*Ecoregions 2017 dataset*
Download Link: https://storage.googleapis.com/teow2016/Ecoregions2017.zip ecoregions.appspot.com



Description: This new map offers a depiction of the 846 ecoregions that represent our living planet. Ecoregions are ecosystems of regional extent. These are color coded on this map to highlight their distribution and the biological diversity they represent. This new map is based on recent advances in biogeography - the science concerning the distribution of plants and animals. The original ecoregions map has been widely used since its introduction in 2001, underpinning the most recent analyses of the effects of global climate change on nature by ecologists to the distribution of the world's beetles to modern conservation planning. In the same vein, our updated ecoregions can now be used to chart progress towards achieving the visionary goal of Nature Needs Half, to protect half of all the land on Earth to save a living terrestrial biosphere.

Citation: Eric Dinerstein, David Olson, Anup Joshi, Carly Vynne, Neil D. Burgess, Eric Wikramanayake, Nathan Hahn, Suzanne Palminteri, Prashant Hedao, Reed Noss, Matt Hansen, Harvey Locke, Erle C Ellis, Benjamin Jones, Charles Victor Barber, Randy Hayes, Cyril Kormos, Vance Martin, Eileen Crist, Wes Sechrest, Lori Price, Jonathan E. M. Baillie, Don Weeden, Kierán Suckling, Crystal Davis, Nigel Sizer, Rebecca Moore, David Thau, Tanya Birch, Peter Potapov, Svetlana Turubanova, Alexandra Tyukavina, Nadia de Souza, Lilian Pintea, José C. Brito, Othman A. Llewellyn, Anthony G. Miller, Annette Patzelt, Shahina A. Ghazanfar, Jonathan Timberlake, Heinz Klöser, Yara Shennan-Farpón, Roeland Kindt, Jens-Peter Barnekow Lillesø, Paulo van Breugel, Lars Graudal, Maianna Voge, Khalaf F. Al-Shammari, Muhammad Saleem, An Ecoregion-Based Approach to Protecting Half the Terrestrial Realm, BioScience, Volume 67, Issue 6, June 2017, Pages 534–545, https://doi.org/10.1093/biosci/bix014

*Global 2010 Tree Cover (30m)*
Download link: https://glad.umd.edu/dataset/global-2010-tree-cover-30-m
Description: Global tree cover data (treecover2010) are per pixel estimates of circa 2010 percent maximum (peak of growing season) tree canopy cover derived from cloud-free annual growing season composite Landsat 7 ETM+ data. A regression tree model estimating per pixel percent tree canopy cover was applied

to annual composites from 2000 to 2012 inclusive (Hansen and others, 2013). Data gaps and noise from individual years were replaced using multi-year median values. First, a median from annual tree canopy cover values from 2009-2011 was used to estimate 2010 tree cover. For pixels still lacking an estimate, the median calculation was expanded to include tree cover values from 2008-2011, then from 2008-2012. Any remaining gaps were filled with tree canopy cover values derived from a regression tree model using all growing season Landsat ETM+ data as inputs. The resulting layer represents estimated maximum tree canopy cover per pixel, 1-100% for the year 2010 in integer values (1-100).
Citation: Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O., and Townshend, J.R.G., 2013, High-Resolution Global Maps of 21st-Century Forest Cover Change: Science, v. 342, no. 6160, p. 850-853. http://www.sciencemag.org/content/342/6160/850.abstract.

*Global Forest Change*
Sharepoint path: Download link in citation
Description: The Global Forest Change Product provides results from time-series analysis of 654,178 Landsat images in characterizing forest extent and change product. For definitions of Forest extent and change refer to Hansen et al., 2013.
Citation: Hansen, M. C., P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend. 2013. "High-Resolution Global Maps of 21st-Century Forest Cover Change." Science 342 (15 November): 850-53. Data available on-line from: http://earthenginepartners.appspot.com/science-2013-global-forest.

*Indonesia Kabupaten*
Sharepoint path: /Walmart Folder for ALL/Component 2/Indonesia_Analysis_GIS_Indonesia/Admin2(Kabupaten)_Indonesia_2020.zip
Description: Kabupaten, second level administrative boundaries for Indonesia provided by CI-IDN.
Citation: Indonesia Geospatial Information Agency. (2020). Indonesia Administrative Boundaries. Available at:  https://tanahair.indonesia.go.id/portal-web

## Section 3: RESTORATION & PROTECTION PRIORITIZATION TOOL

- Colombia ReadMe
- Indonesia ReadMe

# Colombia Coffee: Restoration & Protection Planning Tool README

*This document provides metadata and citations for all datasets used within the Restoration & Protection Planning Tool for Colombia. Links to the tool, the tool scripts, and Earth Engine assets are provided throughout.*

**App Link:** https://ci_external_assets.earthengine.app/view/restoration-and-protection-planning-tool-colombia

**Google Earth Engine Script:**

users/geflanddegradation/CI_code/WMT_Coffee/COL_CofLand_Refor_Priority

**Date Edited:** 12/20/2022

**Authors:** Kristen O'Shea (koshea@conservation.org), David Hunt (dhunt@conservation.org), Anna Ballasiotes(aballasiotes@conservation.org), adopted from work done by Mariano Gonzalez Roglich & Grace StoneCipher

**Language/Software:** Google Earth Engine (GEE) / Javascript

**Script Purpose:** This code creates an application tool for expert users to prioritize areas for restoration in coffee landscapes in Colombia. It contains code for both frontend (user interface) and backend (calculations). The left side of the tool allows the user to select a country and set criteria to calculate and visualize available area, and then to weight different factors to identify high and low priority areas within the available area. After each section, the user has the option to export the layer as a TIF or PNG directly to their computer.

### *Admin Units*

GEE Asset Link:

projects/ci_external_assets/WMT_Coffee/Colombia/2_Intermediate/Colombia_Jurisdictions

Description: The ET Geographic Database contains the limits of the territorial entities, as a result of the demarcation process approved by the competent process of demarcation approved by the competent authorities, fixed or modified by them (departmental (Departmental Assembly, Congress of the Republic of Colombia) and elevated to normativity: Ordinance, Law or Decree,

Law or Decree. They are represented on the cartography of the Instituto Geográfico Agustín Codazzi - IGAC in accordance with Law 14.1. IGAC according to Law 1447 of 2011 and its Regulatory Decree 1170 of 2015. For the Districts the definition and modification of their boundaries is stipulated in Law 1617 of 2013. The Non-Municipalized areas are part of the territorial division, but are not territorial entities (Articles 285 and 286 of the Political Constitution of Colombia, 1991).

Citation: IGAC. 2021. Limite de las entidades Territoriales – Republica de Colombia. Departamentos de Colombia. https://www.colombiaenmapas.gov.co. https://creativecommons.org/licenses/by/4.0/deed.es.

### *Historic Land Cover*

GEE Asset Link:

users/geflanddegradation/toolbox_datasets/lcov_esacc_1992_2018

Description: This dataset is used to define areas for reforestation efforts by looking at pixels that were forest in 1992.

Citation: ESA. Land Cover CCI Product User Guide Version 2. Tech. Rep. (2017). Available at: maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf

### *Current Land Cover*

GEE Asset Link:

projects/ci_external_assets/WMT_Coffee/Colombia/2_Intermediate/Colombia_LULC_active

projects/ci_external_assets/WMT_Coffee/Colombia/2_Intermediate/Colombia_LULC_passive

Description: The landcover map (2018) at scale 1:100,000, represents the visual interpretation of Landsat 8 images for the continental part and sentinel 2 in the insular region. The CORINE Land Cover Methodology adapted for Colombia, scale 1:100,000, was used for its generation (Ideam, IGAC, & Cormagdalena, 2008). The feature scale was 1:100.000 then rasterized at a 30m resolution to align with other datasets. The current landcover layer was used to mask non-restorable landcover categories out of the analysis. Classes included shown below:

| Land Cover Class | Active Restoration | Passive Restoration |
|---|---|---|
| Arbustal abierto | Yes | |
| Arbustal abierto alto | Yes | |
| Arbustal denso | Yes | |
| Areas abiertas sin vegetacion | Yes | Yes |
| Bosque abierto alto | Yes | |
| Bosque abierto bajo | Yes | |
| Bosque de galeria y ripario | Yes | |
| Bosque denso alto | Yes | |
| Bosque denso bajo | Yes | |
| Bosque fragmentado con pastos y cultivos | Yes | Yes |
| Bosque fragmentado con vegetacion secundaria | Yes | Yes |
| Bosque mixto de guandal | Yes | |

| | | |
|---|---|---|
| Caña | Yes | |
| Cafe | Yes | |
| Cultivos permanentes | Yes | |
| Cultivos transitorios | Yes | |
| Herbazal abierto | Yes | |
| Herbazal denso | Yes | |
| Manglar | Yes | |
| Manglar de aguas marinas | Yes | |
| Manglar de aguas mixohalinas | Yes | |
| Mosaico de cultivos con espacios naturales | Yes | Yes |
| Mosaico de cultivos y espacios naturales | Yes | Yes |
| Mosaico de cultivos y pastos | Yes | Yes |
| Mosaico de cultivos, pastos y espacios naturales | Yes | Yes |
| Mosaico de pastos con espacios naturales | Yes | Yes |
| Mosaico de pastos y espacios naturales | Yes | Yes |
| Palma de aceite | Yes | |
| Papa | Yes | |
| Pastos | Yes | |
| Plátano y Banano | Yes | |
| Plantacion forestal | Yes | |
| Turberas | Yes | |
| Vegetacion secundaria | Yes | Yes |

Citation: Instituto de Hidrología, Meteorología y Estudios Ambientales - IDEAM, Instituto Amazónico de Investigaciones Cientificas - SINCHI, Parques Nacionales Naturales de Colombia - PNN. 2021. Mapa de Coberturas de la Tierra Metodología Corine Land Cover. Escala 1:100.000. Periodo 2018.

## *Zoning Data*

GEE Asset Link:

projects/ci_external_assets/WMT_Coffee/Colombia/2_Intermediate/Col_zone_active

projects/ci_external_assets/WMT_Coffee/Colombia/2_Intermediate/Col_zone_passive

Description: To create this layer, we removed all “Áreas con Previa Decisión de Ordenamiento” zones and rasterized the remaining zones at a resolution of 30m. Zones A, B, and C, were included as options in Step 1. The definitions for each zone can be found below

Type A Zones: Zones that guarantee the maintenance of basic ecological processes necessary to ensure the provision of ecosystem services, mainly related to water and climate regulation; assimilation of air and water pollutants; soil formation and protection; protection of unique landscapes and cultural heritage; and biodiversity support.

Type B Zones: Zones that are characterized by having favorable cover for sustainable forest resource management through an integrated forest management approach and integrated management of biodiversity and ecosystem services.

Type C Zones: Zones that due to their biophysical characteristics offer conditions for the development of agroforestry, silvopastoral and other productive activities compatible with the objectives of the forest reserve that must incorporate the forestry component and that do not imply the reduction of the natural forest areas present in their different successional stages.

Citation: Instituto Geográfico Agustín Codazzi, Ministerio de Ambiente y Desarrollo Sostenible (2020) Zonificación Reserva Forestal de Ley 2 de 1959

### *SPAM*

GEE Asset Link:

projects/ci_external_assets/WMT_Coffee/Colombia/2_Intermediate/co_spam_cof_30m

Description: The SPAM Coffee Production layer was used to limit available area for restoration to areas with coffee production. SPAM has production layers for both Arabica and Robusta with an original resolution of 10 km. The extents of both types were combined and exported at a 30m resolution to align with other datasets.

Citation: *SPAM Layers*: International Food Policy Research Institute, 2019, "Global Spatially-Disaggregated Crop Production Statistics Data for 2010 Version 1.1", https://doi.org/10.7910/DVN/PRFF8V, Harvard Dataverse, V3.

### *CIAT Suitability*

GEE Asset Link:

projects/ci_external_assets/WMT_Coffee/Colombia/2_Intermediate/CIAT_Suit_30m

Description: The coffee suitability dataset produced by CIAT was included in Step 1 for users to be able to select if they wanted to view restoration opportunities exclusively within regions suitable for coffee production. The method to produce this dataset consisted of five general steps: initial training data collection, satellite data pre-processing, algorithm training, data classification and results post processing, and fourth, results validation. Validation demonstrated an 80.6% overall accuracy and a user accuracy of 72 and 85% for coffee and non-coffee land use respectively. For additional information on detailed steps of the method, visit the Report: CIAT Coffee Map in 2018 for Colombia Technical Report_EN.docx

Citation: Castro, Fabio and Bunn, Christian; Coffee suitability mapping for coffee in Colombia and Indonesia under past and future conditions; 2021; Alliance of Biodiversity and CIAT; Cali, Colombia.

### *CIAT Coffee Map*

GEE Asset Link:

projects/ci_external_assets/WMT_Coffee/Colombia/2_Intermediate/CIAT_COF_cof_30m

Description: The coffee map dataset produced by CIAT was included in Step 1 for users to be able to select if they wanted to view restoration opportunities exclusively within coffee production areas. The method to produce this dataset consisted of five general steps: initial training data collection, satellite data pre-processing, algorithm training, data classification and results post processing, and fourth, results validation. Validation demonstrated an 80.6% overall accuracy and a user accuracy of 72 and 85% for coffee and non-coffee land use respectively. For additional information on detailed steps of the method, visit the Report: CIAT Coffee Map in 2018 for Colombia Technical Report_EN.docx

Citation: Castro, Fabio and Bunn, Christian; Coffee suitability mapping for coffee in Colombia and Indonesia under past and future conditions; 2021; Alliance of Biodiversity and CIAT; Cali, Colombia.

### *Coffee Suitability Coffea arabica L.*

GEE Asset Link:

projects/ci_external_assets/WMT_Coffee/Colombia/2_Intermediate/UPRA_Coffee_Suitability_30

Description: The Zoning map of suitability for the cultivation of coffee (Coffea arabica L.) in Colombia, at a scale of 1:100.000, is the result of the application of the methodology for the zoning of suitability for commercial crops, described in the document: "Zoning of suitability for the cultivation of coffee (Coffea arabica L.) in Colombia, at a scale of 1:100.000".

Aptitude:

Areas with potential for the establishment and development of the commercial cultivation of coffee (Coffea arabica L.) production, under a legal, normative and technical framework that defines and differentiates them from other possible uses.

Categories:

High Aptitude: Zones with the best conditions from a physical, socioecosystemic and socioeconomic point of view.

Medium Suitability: Areas with moderate physical, socioecosystemic and/or socioeconomic limitations.

Low Suitability: Areas with strong physical, socioecosystemic and/or socioeconomic limitations, which could be adapted with large investments and/or the development of new technologies.

Not suitable: Areas with physical and socio-ecosystemic restrictions that make it impossible to develop the activity.

Legal exclusion: Zones in which, by legal mandate, the development of coffee (Coffea arabica L.) production is not permitted.

Citation: UPRA. 2020. Zonificación de aptitud para el cultivo de café (Coffea arabica L.) en Colombia, a escala 1:100.000. Diciembre 2020. Unidad de Planificacion Rural y Agropecuaria. https://sipra.upra.gov.co/

### *Restoration Priorities*

GEE Asset Link:

projects/ci_external_assets/WMT_Coffee/Colombia/2_Intermediate/RECUPERACION

projects/ci_external_assets/WMT_Coffee/Colombia/2_Intermediate/REHABILITACION

projects/ci_external_assets/WMT_Coffee/Colombia/2_Intermediate/RESTAURACION

Description: Ecological restoration: to restore the degraded ecosystem to a condition similar to the pre-disturbance ecosystem with respect to its composition, structure and functioning. In addition, the resulting ecosystem must be a self-sustaining system and must guarantee the conservation of species, the ecosystem in general and the ecosystem in general as well as most of its goods and services.

Citation: Ministerio de Ambiente y Desarrollo Sostenible. 2015. Grupo de Divulgación de Conocimiento y Cultura Ambiental. Plan Nacional de Restauración: restauración ecológica, rehabilitación y recuperación de áreas disturbadas / Textos: Ospina Arango, Olga Lucia; Vanegas Pinzón, Silvia; Escobar Niño, Gonzalo Alberto; Ramírez, Wilson; Sánchez, John Jairo Bogotá, D.C.: Colombia. 2015. 92 p. ISBN: 978-958-8901-02-2 Medio electrónico o digital

### *Indigenous Peoples and Local Communities*

GEE Asset Link:

projects/ci_external_assets/WMT_Coffee/Colombia/2_Intermediate/IPLCs

Description:

Citation: Instituto Geográfico Agustín Codazzi, Instituto Colombiano de Desarrollo Rural (2019) Mapa de resguardos indígenas

### *Protected Areas*

GEE Asset Link:

projects/ci_external_assets/WMT_Coffee/Colombia/2_Intermediate/Protected_Areas

Description: The Protected Areas layer was used as an option in Step 1 to limit the area available for restoration. The vectors delineating each protected area were imported into Google Earth engine and converted to a raster based on a unique identifier attribute.

Citation: Registro Único Nacional de Áreas Protegidas, Parques Nacionales Naturales de Colombia (2020) Sistema Nacional de Áreas Protegidas de Colombia - SINAP

### *Forests*

GEE Asset Link:

projects/ci_external_assets/WMT_Coffee/Colombia/2_Intermediate/co_dist_norm_30m

Description: The forest layer was used to weight how important it is that restoration occurs on land near existing forests.

Citation: C. Vancutsem, F. Achard, J.-F. Pekel, G. Vieilledent, S. Carboni, D. Simonetti, J. Gallego, L.E.O.C. Aragão, R. Nasi. Long-term (1990-2019) monitoring of forest cover changes in the humid tropics. Science Advances

### *Water Bodies*

GEE Asset Link:

projects/ci_external_assets/WMT_Coffee/Colombia/2_Intermediate/CO_rip_area_30m

Description: The water bodies layer was used to weigh how important it is that restoration occurs within riparian areas.

Citation: Instituto Geográfico Agustín Codazzi - Subdirección de Geografía y Cartografía– Grupo Interno de Trabajo Generación de Datos y Productos Cartográficos (2019) Cartografía Básica Digital Integrada. República de Colombia. Escala 1:100.000.

### *Population*

GEE Asset Link:

projects/ci_external_assets/WMT_Coffee/Colombia/2_Intermediate/high_population_mask

Description: For the estimation of population density, we used data from the 2018 census of the National Administrative Department of Statistics (DANE), divided the number of habitants in the reported area and hectares, and selected those areas with more than 150 habitants/ha, as the most density populated areas in urban areas.

Citation: Departamento Administrativo Nacional de Estadistica. DANE. 2018. Censo Nacional de Poblacion y Vivienda. CNPV – 2018. Marco Geoestadistico Nacional. Direccion de Censos y Demografía. 38 p.

### *Slope*

GEE Asset Link:

USGS/SRTMGL1_003

Description: The slope layer was used to weight how important it is that restoration occurs on land with high slope. Elevation data was imported from the Shuttle Radar Topography Mission (SRTM). Then a slope layer was derived using the ee.Terrain.slope function in Google Earth Engine and clipped to Colombia. Then the layer was resampled to prioritize higher slopes based on input from the Colombia field team. Finally, it was normalized nationally.

Citation: Farr, T. G., et al. (2007), The Shuttle Radar Topography Mission, Rev. Geophys., 45, RG2004, doi:10.1029/2005RG000183.

| Slope Range | Reclassified Value |
|---|---|
| 0 - 3 degrees | 5 |
| 3– 8 degrees | 4 |
| 9 – 15 degrees | 3 |
| 15 – 30 degrees | 2 |
| 30+ degrees | 1 |

### *Carbon Sequestration Potential*

GEE Asset Link:

projects/ci_external_assets/WMT_Coffee/Colombia/2_Intermediate/COL_ABG_norm

Description: This layer utilizes the Cook-Patton et al. dataset that shows the rate at which forests could capture carbon form the atmosphere and store it in aboveground live biomass over the next 30 years. To prepare this dataset, the Cook-Patton et al. data was clipped to the national Colombia boundary and normalized so all values were between 0 and 1, where 0 was the lowest priority and 1 was the highest priority.

Citation: Cook-Patton, S.C., Leavitt, S.M., Gibbs, D. et al. Mapping carbon accumulation potential from global natural forest regrowth. Nature 585, 545–550 (2020). https://doi.org/10.1038/s41586-020-2686-x

### *Biodiversity*

GEE Asset Link:

projects/ci_external_assets/WMT_Coffee/Colombia/2_Intermediate/co_rsr_30m_norm

Description: The biodiversity layer was used to weight how important it is that restoration occurs on land with high range rarity.

Citation: Derived by CI from: The IUCN Red List of Threatened Species. Version 2021-2. https://www.iucnredlist.org, BirdLife International and Handbook of the Birds of the World (2020) Bird species distribution maps of the world. Version 2020.1. Available at http://datazone.birdlife.org/species/requestdis [obtained 6/1/2020].

### *Fire*

GEE Asset Link:

projects/ci_external_assets/WMT_Coffee/Colombia/2_Intermediate/Fire_density_normalized

Description: All MODIS fire detection points for Indonesia from between 01/01/2008 and 12/31/2018 were downloaded from the Fire Information for Resource Management System (FIRMS) Archive Download. These points were filtered to remove all fires with a less than 30 percent confidence rating. Then, a kernel density was run in ArcGIS Pro to create a density raster for the locations of the historical fire detection points. Finally, this density raster was normalized between 0 and 1 and the values were inverted so areas with the value of 0 represent the highest density, and therefore, the lowest priority for restoration.

Citations: MODIS Collection 61 NRT Hotspot / Active Fire Detections MCD14DL distributed from NASA FIRMS.Available on-line [https://earthdata.nasa.gov/firms]. 10.5067/FIRMS/MODIS/MCD14DL.NRT.0061

### *Watersheds*

GEE Asset Link:

projects/ci_external_assets/WMT_Coffee/Colombia/2_Intermediate/watershed_headwaters

Description: The watershed dataset was included to prioritize valuable watershed headwaters for restoration. Using the DEM from SRTM (described the Slope section above), the top 20 percentile of elevation areas were calculated for each watershed using zonal statistics. These regions were then extracted for each watershed and reclassified to 1, while the areas outside the headwaters were reclassified to 0.

Citation: IDEAM. 2010. Mapa de Zonificación Hidrográfica. Colombia. Escala 1:500.000. Año 2010.

### *Roads*

GEE Asset Link:

projects/ci_external_assets/WMT_Coffee/Colombia/2_Intermediate/Roads_dist_passive projects/ci_external_assets/WMT_Coffee/Colombia/2_Intermediate/Roads_mask_active

Description: The roads dataset was incorporated to address the issue of accessibility to a given restoration site. For active restoration, all areas outside a buffer of 2.5 kilometers from a road were removed from the analysis because these sites would be too inaccessible and planting activities too challenging. All sites within the 2.5-kilometer distance were determined to be equally suitable for restoration.

For passive restoration, all areas within the 2.5-kilometer buffer were given values of zero because they would be in close proximity to areas that could potentially interrupt natural regeneration. We then created a distance raster from this 2.5-kilometer buffer and normalized the values so areas with the largest distance from the buffer had a value of 1 and areas next to the buffer approached values of 0. This incentivizes passive restoration in areas that are least accessible to humans.

| Road Type | Description | Active Restoration | Passive Restoration |
|---|---|---|---|
| 1 | Paved roadway, which is a smooth, hard and durable layer of asphalt, cement, pavers or other of asphalt, cement, cobblestones or other materials that cover the ground to make it firm and level. the ground is covered to make it firm and level. | Yes | Yes |
| 2 | Roads without pavement, however, it is a smooth road with affirmed so that it is passable all the year round. | Yes | Yes |
| 3 | Roads that are not paved, also lack pavement, which makes them only passable in dry weather. only passable in dry weather. | Yes | Yes |
| 4 | Road under construction, usually in continuity with or connecting to a paved road. | Yes | Yes |
| 5 | Unpaved, passable in dry weather | Yes | Yes |
| 6 | No pavement passable in dry weather | Yes | Yes |
| 7 - Camino-sendero | Road or path of road adaptation, generally rural. generally rural, along which mainly pedestrians and mainly pedestrians and animals. animals. The surface is not paved or paved or paved. | | |
| 8 - Peatonal-urbana | Narrow road in urban areas, that has been designed for foot traffic. of people. The surface is paved or or paved. | | |

Citation: Instituto Geografico Agustin Codazzi - IGAC. 2019. Base de datos vectorial basica. Colombia. Escala 1:100.000. Año 2019. https://www.igac.gov.co/sites/igac.gov.co/files/igac_co_cartografiabasica_v2.2.pdf

## Coffee Coincident Deforestation

GEE Asset Link:

projects/ci_external_assets/WMT_Coffee/Colombia/2_Intermediate/prioritized_municipios_q2_30

Description: This metric is just the portion of a biotic unit's deforestation coincident with coffee aggregated to the municipal level.

Citation: Follett, Forrest, Christy Melhart Slay (2022) "Comparison of Administrative Units' Contribution of Commodity-associated Tree Cover Loss to Priority Forested Ecoregions." The Sustainability Consortium. https://sustainabilityconsortium.org/download/follett-slay-2022-comparison/

## Reforestation Score

GEE Asset Link:

projects/ci_external_assets/WMT_Coffee/Colombia/2_Intermediate/prioritized_municipios_reforestat

Description: This metric looks at the portion of a biotic unit's non-natural land cover that is coffee, weighted by the density of endemic species in the biotic unit, and aggregated to the municipal level. This metric is meant to indicate municipios where reforesting coffee would have the biggest impact on biodiversity.

Citation: Follett, Forrest, Christy Melhart Slay (2022) "Comparison of Administrative Units' Contribution of Commodity-associated Tree Cover Loss to Priority Forested Ecoregions." The Sustainability Consortium. https://sustainabilityconsortium.org/download/follett-slay-2022-comparison/

## Forests At Risk

GEE Asset Link:

projects/forestatrisk/assets/v1_2020/fcc_123

projects/forestatrisk/assets/v1_2020/fcc_2050

Description: The Forest at Risk dataset includes spatial models of deforestation in 92 countries covering all the tropical moist forests in the world. Rasters have a resolution of 30 m and are available in Albers equal area ("aea") conic projections (one different projection for each continent).

Citation: Vieilledent G., C. Vancutsem, C. Bourgoin, P. Ploton, P. Verley, and F. Achard. 2022. Spatial scenario of tropical deforestation and carbon emissions for the 21st century. bioRxiv. doi: 10.1101/2022.03.22.485306

### Irrecoverable Carbon

GEE Asset Link:

projects/ci_geospatial_assets/CIE/Final_Assets/Irrecoverable_Carbon_Total_2018_30m

Description:

Irrecoverable Carbon (t/ha) refers to the vast stores of carbon in nature that are vulnerable to release from human activity and, if lost, could not be restored by 2050 — when the world must reach net-zero emissions to avoid the worst impacts of climate change.  This specific dataset is 30m spatial resolution.

Citation:

Noon, M.L., Goldstein, A., Ledezma, J.C. et al. Mapping the irrecoverable carbon in Earth's ecosystems. Nat Sustain 5, 37–46 (2022). https://doi.org/10.1038/s41893-021-00803-6

### Directions for making script updates:

Adding a New Restoration Area Mask:

- Step 1: Bring New Mask Dataset into Landcover Selector Widget
- Step 2. Reclassify as 1's and 0's , with 1's areas you want to keep and 0's areas you do not want to keep
- Step 3. Multiply restorationLandcover (active or passive) variable by new mask
- Adding New Restoration Priority Weighted Layer:
- Step 1: Generate a New Slider
- Copy Section 14 and Update Section Name
- Update Variable Names
- Update Palette
- Add Dataset toDisplay Layer Button Function
- Add Slider to Panel
- Step 2: Bring in normalized national dataset you wish to weight
- Add variable for new dataset in Run Weighted Overlay section
- Step 3: Calculate the weighted layer
- Create *newvariable*_weighted in the Run Analysis Button function under the "weight each layer using user-input values" comment
- Step 4: Add new weighted layer to the rest
- Add *newvariable*_weighted under the "add weighted layers together" comment

# Indonesia Coffee: Restoration & Protection Planning Tool README

*This document provides metadata and citations for all datasets used within the Restoration & Protection Planning Tool for Indonesia. Links to the tool, the tool scripts, and Earth Engine assets are provided throughout.*

**Script Google Earth Engine Location:**
users/geflanddegradation/CI_code/WMT_Coffee/IDN_CofLand_Refor_Priority

**App Link:** [https://ci-external-assets.earthengine.app/view/restoration-and-protection-planning-tool-indonesia](https://ci-external-assets.earthengine.app/view/restoration-and-protection-planning-tool-indonesia)

**Date Updated:** 12/20/2022

**Authors:** Kristen O'Shea (koshea@conservation.org), David Hunt (dhunt@conservation.org), Anna Ballasiotes (aballasiotes@conservation.org), adopted from work done by Mariano Gonzalez Roglich & Grace StoneCipher

**Privacy:** This tool's use is currently restricted to project partners.

**Language/Software:** Google Earth Engine (GEE)

**Script Purpose:** This code creates an application tool for restoration expert users to prioritize areas for restoration in coffee landscapes in Indonesia. It contains code for both frontend (user interface) and backend (calculations). The left side of the tool allows the user to select a country and set criteria to calculate and visualize available area, and then to weight different factors to identify high and low priority areas within the available area. After each section, the user has the option to export the layer as a raster to their Google Drive.

## *Admin Units*

GEE Asset Link: projects/ci_geospatial_assets/admin_units/gadm36_adm1_100m

Description: This dataset was used to delineate boundaries for the areas the user wants to explore. It was filtered to only include Indonesia jurisdictions.

Citation: Global Administrative Areas (2012). GADM database of Global Administrative Areas, version 2.0. [online] URL: [www.gadm.org](http://www.gadm.org).

### *Historic Land Cover*

GEE Asset Link: users/geflanddegradation/toolbox_datasets/lcov_esacc_1992_2018

Description: This dataset is used to define areas for reforestation efforts by looking at pixels that were forest in 1992.

Citation: ESA. Land Cover CCI Product User Guide Version 2. Tech. Rep. (2017). Available at: maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf

### *Current Land Cover*

GEE Asset Links:
projects/ci_external_assets/WMT_Coffee/Indonesia/2_Intermediate/Indonesia_landcover_2019_activ
projects/ci_external_assets/WMT_Coffee/Indonesia/2_Intermediate/Indonesia_landcover_2019_passive

Description: This dataset is landcover data based on the Ministry of Environment and Forestry of the Republic of Indonesia published in 2020 from interpretation of LDCM (The Landsat Data Continuity Mission)/Landsat 8 OLI. The feature scale was 1:250.000, which was then rasterized to 30m resolution to align with other datasets. The current landcover layer was used to mask non-restorable landcover categories out of the analysis. Classes included shown below:

| Land Cover Class | Active Restoration | Passive Restoration |
|---|---|---|
| Bare Land (Land Clearing/Dirt) | Yes | |
| Bush / Shrub | Yes | Yes |
| Dryland Agriculture | Yes | Yes |
| Estate Crop Plantation | Yes | Yes |
| Plantation Forest | Yes | |
| Secondary Dry Land Forest | Yes | Yes |
| Secondary Mangrove Forest | Yes | Yes |
| Secondary Swamp Forest | Yes | Yes |
| Shrub-Mixed Dryland Farm | Yes | |
| Swamp Shrub | Yes | Yes |

Citation: Ministry of Environment and Forestry (MoEF) of Indonesia, 2019

### *SPAM*

GEE Asset Link:
projects/ci_external_assets/WMT_Coffee/Indonesia/2_Intermediate/spam_cof_30m

Description: The SPAM Coffee Production layer was used to limit available area for restoration to areas with coffee production. SPAM has production layers for both Arabica and Robusta with an original resolution of 10 km. The extents of both types were combined and exported at a 30m resolution to align with other datasets.

Citation: *SPAM Layers*: International Food Policy Research Institute, 2019, "Global Spatially-Disaggregated Crop Production Statistics Data for 2010 Version 1.1", https://doi.org/10.7910/DVN/PRFF8V, Harvard Dataverse, V3.

### *CIAT Suitability*

GEE Asset Link:
projects/ci_external_assets/WMT_Coffee/Indonesia/2_Intermediate/Indonesia_Coffee_Suitability_30m

Description: The coffee suitability dataset produced by CIAT was included in Step 1 for users to be able to select if they wanted to view restoration opportunities exclusively within regions suitable for coffee production. The method to produce this dataset consisted of five general steps: initial training data collection, satellite data pre-processing, algorithm training, data classification and results post processing, and fourth, results validation. Validation demonstrated an 85.4% overall accuracy and a user accuracy of 72.2 and 92.1% for coffee and non-coffee land use respectively. For additional information on detailed steps of the method, visit the Report: CIAT Coffee Map in 2018 for Indonesia Technical Report_EN.docx

Citation: Castro, Fabio and Bunn, Christian; Coffee suitability mapping for coffee in Colombia and Indonesia under past and future conditions; 2021; Alliance of Biodiversity and CIAT; Cali, Colombia.

### *CIAT Coffee Map*

GEE Asset Link:
projects/ci_external_assets/WMT_Coffee/Indonesia/2_Intermediate/Indonesia_Coffee_30m

Description: The coffee map dataset produced by CIAT was included in Step 1 for users to be able to select if they wanted to view restoration opportunities exclusively within coffee production areas. The method to produce this dataset consisted of five general steps: initial training data collection, satellite data pre-processing, algorithm training, data classification and results post processing, and fourth, results validation. Validation demonstrated an 85.4% overall accuracy and a user accuracy of 72.2 and 92.1% for coffee and non-coffee land use respectively. For additional information on detailed steps of the method, visit the Report: CIAT Coffee Map in 2018 for Indonesia Technical Report_EN.docx

Citation: Castro, Fabio and Bunn, Christian; Coffee suitability mapping for coffee in Colombia and Indonesia under past and future conditions; 2021; Alliance of Biodiversity and CIAT; Cali, Colombia.

### *Protected Areas*

GEE Asset Link: WCMC/WDPA/current/polygons

Description: The Protected Areas layer was used as an option to limit the area available for restoration. The vectors delineating each protected area were imported into Google Earth engine and converted to a raster based on a unique identifier attribute.

Citation: UNEP-WCMC and IUCN (2021), Protected Planet: The World Database on Protected Areas (WDPA) and World Database on Other Effective Area-based Conservation Measures (WD-OECM) [Online], September 2021, Cambridge, UK: UNEP-WCMC and IUCN. Available at: www.protectedplanet.net.

### *Forests*

GEE Asset Link:
projects/ci_external_assets/WMT_Coffee/Indonesia/2_Intermediate/idn_distance_norm_30m

Description: The forest layer was used to weight how important it is that restoration occurs on land near existing forests. Tiles of the 30m Tropical Moist Forest transition map covering Indonesia were downloaded into ArcMap. Then forest classes were selected and reclassified as 1. Non-forest classes were classified as 0. Then a distance from forest layer was generated and uploaded into Google Earth Engine.

Citation: C. Vancutsem, F. Achard, J.-F. Pekel, G. Vieilledent, S. Carboni, D. Simonetti, J. Gallego, L.E.O.C. Aragão, R. Nasi. Long-term (1990-2019) monitoring of forest cover changes in the humid tropics. Science Advances

### *Peatlands*

GEE Asset Link:
projects/ci_external_assets/WMT_Coffee/Indonesia/2_Intermediate/Indonesia_peat30m

Description: The peat layer was used to weight how important it is that restoration occurs in areas that are peatland. The binary layer came from PEATMAP and was resampled to 30m.

Citation: Xu, Jiren, Paul Morris, Junguo Liu, and Joseph Holden. “PEATMAP: Refining Estimates of Global Peatland Distribution Based on a Meta-Analysis.” Catena 160 (January 2018): 134–40. https://doi.org/10.1016/j.catena.2017.09.010.

### *Water Bodies*

GEE Asset Links:
projects/ci_external_assets/WMT_Coffee/Indonesia/2_Intermediate/Water_mask
projects/ci_external_assets/WMT_Coffee/Indonesia/2_Intermediate/Riparian_zone

Description: The water bodies layer was used to both mask water surfaces out of the analysis and to weight the importance of riparian areas for restoration potential. The water mask was created by merging the water bodies (lakes and rivers) into one layer and converting to a raster with a value of 1 and resolution of 30m. These areas can be removed from analysis since restoration cannot be performed on water surfaces. (These areas were most likely also excluded by the landcover masks, but this is an extra precaution.)

The riparian areas were created by buffering different water body types. Small rivers, less than 30m wide, and lakes were buffered by 50m, while large rivers, over 30m wide, were buffered by 100m. These buffer layers were merged into one layer and converted into a raster where riparian zones had a value of 1 and non-riparian zones had a value of 0. This binary raster layer ensures that riparian zones are prioritized in the restoration analysis, while non-riparian zones are not prioritized, but not excluded from the analysis.

Citation: Indonesia Geospatial Information Agency, 2019.

### *Impervious Surfaces*

GEE Asset Links:
projects/ci_external_assets/WMT_Coffee/Indonesia/2_Intermediate/Impervious_surface_distance_active
projects/ci_external_assets/WMT_Coffee/Indonesia/2_Intermediate/Impervious_surface_distance_passive
projects/ci_external_assets/WMT_Coffee/Indonesia/2_Intermediate/Impervious_surface_buffer_mask

Description: The 30m resolution impervious surfaces layer was used to both mask impervious surfaces out of the analysis and to weight the importance of the distance to impervious surfaces for restoration potential. The impervious surface mask was created by buffering impervious surfaces by 30m and adding the buffer to the impervious surface areas. Then, a binary layer was created where impervious surfaces had a value of 1. This allowed all impervious surfaces and all areas immediately adjacent to impervious surfaces to be excluded from the analysis.
The distance raster was created by running a Euclidean distance on the impervious surface plus buffer layer and normalizing the resulting raster to between 0 and 1, resulting in impervious surfaces having a value of 0 and areas with the maximum distance having a value of 1. To include in the tool, the values in the normalized distance raster were inverted to create a second distance raster where the impervious surfaces had a value 1 and areas with the maximum distance had a value of 0. The non-inverted raster was used in the passive restoration analysis to incentivize areas far from built-up regions, and the inverted raster was used for the active restoration analysis to incentivize areas close, and therefore more accessible, to urban areas.

Citation: Zhang, X., Liu, L., Wu, C., Chen, X., Gao, Y., Xie, S., and Zhang, B.: Development of a global 30m impervious surface map using multisource and multitemporal remote sensing datasets with the Google Earth Engine platform, Earth Syst. Sci. Data, 12, 1625–1648, https://doi.org/10.5194/essd-12-1625-2020, 2020.

### *Slope*

GEE Asset Link: USGS/SRTMGL1_003

Description: The slope layer was used to weight how important it is that restoration occurs on land with high slope. Elevation data was imported from the Shuttle Radar Topography Mission (SRTM). Then a slope layer was derived using the ee.Terrain.slope function in Google Earth Engine and clipped to Indonesia. Then the layer was resampled to prioritize higher slopes based on input from the Indonesia field team. Finally, it was normalized nationally.

Citation: Farr, T. G., et al. (2007), The Shuttle Radar Topography Mission, Rev. Geophys., 45, RG2004, doi:10.1029/2005RG000183.

| Slope Range | Reclassified Value |
| --- | --- |
| 0 - 8 degrees | 1 |
| 9 – 15 degrees | 2 |
| 16 – 25 degrees | 3 |
| 26 – 40 degrees | 4 |
| 40+ degrees | 5 |

### *Carbon Sequestration Potential*

GEE Asset Link:
projects/ci_external_assets/WMT_Coffee/Indonesia/2_Intermediate/IDN_ABG_norm

Description: This layer utilizes the Cook-Patton et al. dataset that shows the rate at which forests could capture carbon form the atmosphere and store it in aboveground live biomass over the next 30 years. To prepare this dataset, the Cook-Patton et al. data was clipped to the national Indonesian boundary and normalized so all values were between 0 and 1, where 0 was the lowest priority and 1 was the highest priority.

Citation: Cook-Patton, S.C., Leavitt, S.M., Gibbs, D. et al. Mapping carbon accumulation potential from global natural forest regrowth. Nature 585, 545–550 (2020). https://doi.org/10.1038/s41586-020-2686-x

### *Biodiversity*

GEE Asset Link: projects/ci_external_assets/WMT_Coffee/Indonesia/2_Intermediate/rsr_30m

Description: The range rarity layer was used to weight how important it is that restoration occurs on land with high range rarity. It was resampled to a 30m resolution.

Citation: Derived by CI from: The IUCN Red List of Threatened Species. Version 2021-2. https://www.iucnredlist.org, BirdLife International and Handbook of the Birds of the World (2020) Bird species distribution maps of the world. Version 2020.1. Available at http://datazone.birdlife.org/species/requestdis [obtained 6/1/2020].

### *Fire*

GEE Asset Link:
projects/ci_external_assets/WMT_Coffee/Indonesia/2_Intermediate/Fire_density_normalized

Description: All MODIS fire detection points for Indonesia from between 01/01/2008 and 12/31/2018 were downloaded from the Fire Information for Resource Management System (FIRMS) Archive Download. These points were filtered to remove all fires with a less than 30 percent confidence rating. Then, a kernel density was run in ArcGIS Pro to create a density raster for the locations of the historical fire detection points. Finally, this density raster was normalized between 0 and 1 and the values were inverted so areas with the value of 0 represent the highest density, and therefore, the lowest priority for restoration.

Citations: MODIS Collection 61 NRT Hotspot / Active Fire Detections MCD14DL distributed from NASA FIRMS.Available on-line [https://earthdata.nasa.gov/firms].
10.5067/FIRMS/MODIS/MCD14DL.NRT.0061

### *Watersheds*

GEE Asset Link:
projects/ci_external_assets/WMT_Coffee/Indonesia/2_Intermediate/watershed_headwaters
Description: The watershed dataset was included to prioritize valuable watershed headwaters for restoration. Using the DEM from SRTM (described the *Slope* section above), the top 20 percentile of elevation areas were calculated for each watershed using zonal statistics. These regions were then extracted for each watershed and reclassified to 1, while the areas outside the headwaters were reclassified to 0.

Citation: Ministry of Environment and Forestry, 2019

### *Zones*

GEE Asset Link:
projects/ci_external_assets/WMT_Coffee/Indonesia/2_Intermediate/Indonesia_Zones_30m

Description: The dataset of Ministry of Environment and Forestry zones was included in Step 1 for users to be able to select if they wanted to view restoration opportunities exclusively within relevant zones. The zones included are in the table below. This layer was converted to a raster with a 30m resolution to align with other datasets in the restoration tool.

| Zone |
| --- |
| Conservation Areas |
| Marine Conservation Areas |
| Other Non-Forest Use |
| Production Forest |
| Production Forest - Convertible |
| Production Forest - Limited |
| Protection Forest |
| Waterbody |

Citation: Ministry of Environment and Forestry, 2020

### *Social Forestry*

GEE Asset Link:
projects/ci_external_assets/WMT_Coffee/Indonesia/2_Intermediate/Social_Forestry_Designated_30m
projects/ci_external_assets/WMT_Coffee/Indonesia/2_Intermediate/Social_Forestry_Indicative_30m

Description: Two different social forestry layers were included in Step 1 for users to be able to select if they wanted to view restoration opportunities exclusively within social forestry boundaries. The first layer was designated social forestry areas, which are areas approved and social forestry and may be ready to implement restoration actions imminently. The second layer was indicative social forestry areas, which shows potential areas where social forestry could be designated, but local stakeholders have not applied for a legal permit to use the area. These two layers were both converted to rasters with a 30m resolution to align with other datasets in the restoration tool.

Citation: Ministry of Environment and Forestry, 2021

### *Coffee Coincident Deforestation*

GEE Asset Link:
projects/ci_external_assets/WMT_Coffee/Colombia/2_Intermediate/prioritized_municipios_q2_30

Description: This metric is the portion of a biotic unit's deforestation coincident with coffee aggregated to the administrative 2 level.

Citation: Follett, Forrest, Christy Melhart Slay (2022) "Comparison of Administrative Units' Contribution of Commodity-associated Tree Cover Loss to Priority Forested Ecoregions." The Sustainability Consortium. https://sustainabilityconsortium.org/download/follett-slay-2022-comparison/

### *Reforestation Score*

GEE Asset Link: projects/ci_external_assets/WMT_Coffee/Indonesia/2_Intermediate/Indonesia_Score_30m

Description: This metric looks at the portion of a biotic unit's non-natural landcover that is coffee, weighted by the density of endemic species in the biotic unit, and aggregated to the municipal level. This metric is meant to indicate jurisdictions where reforesting coffee would have the biggest impact on biodiversity.

Citation: Follett, Forrest, Christy Melhart Slay (2022) "Comparison of Administrative Units' Contribution of Commodity-associated Tree Cover Loss to Priority Forested Ecoregions." The Sustainability Consortium. https://sustainabilityconsortium.org/download/follett-slay-2022-comparison/

### *Forests At Risk*

GEE Asset Link:
projects/forestatrisk/assets/v1_2020/fcc_123
projects/forestatrisk/assets/v1_2020/fcc_2050

Description: The Forest at Risk dataset includes spatial models of deforestation in 92 countries covering all the tropical moist forests in the world. Rasters have a resolution of 30 m and are available in Albers equal area ("aea") conic projections (one different projection for each continent).

Citation: Vieilledent G., C. Vancutsem, C. Bourgoin, P. Ploton, P. Verley, and F. Achard. 2022. Spatial scenario of tropical deforestation and carbon emissions for the 21st century. bioRxiv. doi: 10.1101/2022.03.22.485306

### *Irrecoverable Carbon*

GEE Asset Link:
projects/ci_geospatial_assets/CIE/Final_Assets/Irrecoverable_Carbon_Total_2018_30m

Description:
Irrecoverable Carbon (t/ha) refers to the vast stores of carbon in nature that are vulnerable to release from human activity and, if lost, could not be restored by 2050 — when the world must reach net-zero emissions to avoid the worst impacts of climate change. This specific dataset is 30m spatial resolution.

Citation:
Noon, M.L., Goldstein, A., Ledezma, J.C. *et al.* Mapping the irrecoverable carbon in Earth's ecosystems. *Nat Sustain* **5**, 37–46 (2022). https://doi.org/10.1038/s41893-021-00803-6

### *Directions for making script updates:*

*Adding a New Restoration Area Mask:*

- Step 1: Navigate to the Landcover Selector Widget Section
- Step 2: Import the New Mask as an Image and add it to the template layer.
- Step 3: Reclassify the mask so areas you want to keep are 1s and areas you want to mask out are 0s.
- Step 4: Multiply the mask by the restorationLandcover variable and update the mask to values greater than 0.

Example: Lines 175-179

*Adding New Restoration Priority Weighted Layer:*

- Step 1: Generate a New Slider
    - Copy and paste a previous slider code section at the end of the slider sections
    - Update the section variable for the new dataset you are adding as a weight
    - Update the Labels
    - Add the new panel defined at the end of the section to the Build Priorities Panel
- Step 2: Bring in normalized national dataset you wish to weight
- Step 3: Create a variable to calculate the weighted layer
- Step 4: Add new weighted layer to the rest to get the new restorationPriority_unscaled