UNIVERSIDAD
NACIONAL
DE COLOMBIA

Coca Crop Classification and Mapping using Spectral, Temporal and Spatial Features from Satellite Imagery for the Catatumbo region in Colombia - 2019

Camilo Andrés Albarracín Barrera

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Estadística
Bogotá D.C., Colombia
2023

Coca Crop Classification and Mapping using Spectral, Temporal and Spatial Features from Satellite Imagery for the Catatumbo region in Colombia - 2019

Camilo Andrés Albarracín Barrera

Trabajo de grado presentado como requisito parcial para optar al título de:
Magíster en Ciencias - Estadística

Director(a):
Ph.D. Martha Patricia Bohorquez Castañeda

Línea de Investigación:
Estadística Espacial
Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Estadística
Bogotá D.C., Colombia
2023

Agradecimientos

En primer lugar, me gustaría expresar mi más profundo agradecimiento a mi directora de tesis, Martha Patricia Bohorquez Castañeda, por su constante e inquebrantable apoyo, paciencia y valiosa orientación a lo largo del desarrollo de esta investigación. Su amplio conocimiento y experiencia en el campo de la modelización espacio-temporal han sido fundamentales para la culminación exitosa de este proyecto.

Asimismo, quisiera agradecer a mis colegas del Sistema Integrado de Monitoreo de Cultivos Ilícitos (SIMCI), no solo por facilitar la información para el desarrollo de este trabajo, sino por las valiosas discusiones técnicas que sostuvimos alrededor de los sensores remotos y las dinámicas del cultivo den el paías, la cuales permitieron enriquecer el contenido y calidad de este trabajo. Su compromiso y dedicación han sido una fuente de inspiración y motivación para mí.

Por último, pero no menos importante, me gustaría agradecer a mi familia: a mi madre, Flor Alba, a mi padre David, a mis hermanos Juan David y Gadiel y a mi tía Elsa por su amor y apoyo en cada etapa de mi vida académica y personal. Sin su aliento y fe en mí, no habría sido posible alcanzar este logro.

Resumen

Título: Clasificación y mapeo de cultivos de coca utilizando características espectrales, temporales y espaciales a partir de imágenes satelitales para la región del Catatumbo en Colombia - 2019

El monitoreo de cultivos de coca es esencial para la formulación de políticas públicas de drogas a nivel global, especialmente con la expansión hacia a países no tradicionales. Como el principal productor de cocaína del mundo, Colombia ejemplifica los desafíos inherentes al monitoreo de este cultivo. El modelo actual de monitoreo, establecido en colaboración con la Oficina de las Naciones Unidas contra la Droga y el Delito (UNODC), ofrece una estimación robusta pero sujeta a mejoras en términos de oportunidad y eficiencia, dado que depende de la interpretación visual de imágenes satelitales anuales. Este trabajo presenta una metodología innovadora que emplea XGBoost con datos multiespectrales y espacio-temporales, principalmente de imágenes Sentinel-2. El flujo de trabajo escalable utiliza Google Earth Engine (GEE) para acceder a las imágenes satelitales y extraer variables para la clasificación. Los modelos XGBoost se entrenan para diferenciar entre coca y no coca y se optimizan utilizando un método de validación cruzada espacial. Al aplicarse en dos zonas de Putumayo, Colombia, esta metodología produce una puntuación Kappa de 0,7512 usando datos de Sentinel-2, superando la puntuación Kappa de 0,7090 alcanzada en trabajos anteriores. Este avance representa un paso significativo en la precisión de la clasificación a gran escala de cultivos de coca. Un experimento complementario utilizando imagenes de Planet, de mayor resolución, en una de las zonas para 2021 produjo una precisión menor pero una mejor delimitación geométrica, verificada al evaluar la homogeneidad espectral entre polígonos clasificados y polígonos de referencia. Esta notable mejora en las metodologías de clasificación de cultivos tiene el potencial de fortalecer las operaciones de las fuerzas del orden, perfeccionar las políticas de drogas e influir en las relaciones internacionales.

Palabras clave: clasificación de cultivos, coca, Colombia, XGBoost, datos multiespectrales, espacio-temporal, Google Earth Engine, Sentinel-2, Planet, política de drogas.

Abstract

Title: Coca Crop Classification and Mapping using Spectral, Temporal and Spatial Features from Satellite Imagery for the Catatumbo region in Colombia - 2019

Coca crop monitoring is essential for the formulation of public drug policies at the global level, especially with the expansion into non-traditional countries. As the world's leading cocaine producer, Colombia exemplifies the challenges inherent in monitoring this crop. The current monitoring model, established in collaboration with the United Nations Office on Drugs and Crime (UNODC), offers a robust estimate but is subject to improvement in terms of timeliness and efficiency, as it relies on visual interpretation of annual satellite imagery. This paper presents an innovative methodology that employs XGBoost with multispectral and spatio-temporal data, mainly from Sentinel-2 imagery. The scalable workflow uses Google Earth Engine (GEE) to access satellite imagery and extract variables for classification. The XGBoost models are trained to differentiate between coca and non-coca and optimized using a spatial cross-validation method. When applied in two areas of Putumayo, Colombia, this methodology produced a Kappa score of 0,7512 using Sentinel-2 data, surpassing the Kappa score of 0,7090 achieved in previous work. This advance represents a significant step forward in the accuracy of large-scale coca field classification. A complementary experiment using higher resolution Planet imagery in one of the zones for 2021 produced a lower accuracy but better geometric delineation, verified by assessing spectral homogeneity between classified polygons and reference polygons. This marked improvement in crop classification methodologies has the potential to strengthen law enforcement operations, refine drug policies and influence international relations.

Keywords: crop classification, coca, Colombia, XGBoost, multispectral data, spatial-temporal, Google Earth Engine, Sentinel-2, Planet, drug policy.

Content

List of Figures

List of Tables

1 Introduction

The persistent and evolving challenges posed by the illicit cultivation and production of coca leaves and their subsequent transformation into cocaine have burdened many nations over the past century. In particular, Colombia, Peru, and Bolivia have experienced significant socio-economic and environmental impacts due to the pervasive presence of this illicit industry [UNODC, 2023]. Increasingly, however, the problem of illicit coca cultivation is extending beyond these traditional coca-growing countries. A range of Latin American and Central American nations are becoming entwined in this illicit activity, grappling with the complexities of the international drug trade and regional disparities in law enforcement capacities [InfoBAE, 2022, El Financiero, 2023, El Tiempo, 2023]. This issue underscores the urgent need for a universally accessible, affordable, and effective systems to monitor coca cultivation and comprehend its extent beyond the traditional coca-growing regions. Figure **1-1** illustrates the potential regions for coca highlighting that any region between the two tropics and below 1.800 meters above sea level has the potential for coca cultivation, including countries from Africa, Asia and Oceania.

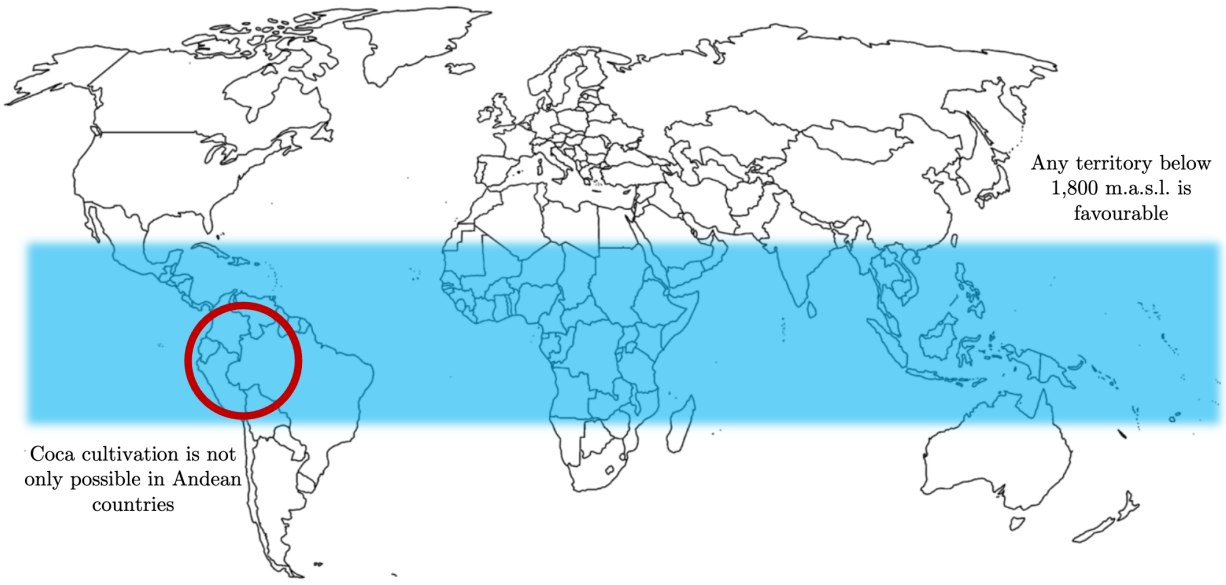

Figure 1-1: Potential regions for coca cultivation
Source: Author's own

For the past decade, remote sensing (RS) technology has emerged as a critical tool for generating agricultural statistics. With advancements in technology and increased data accessibility, this method leverages a combination of resources, such as the electromagnetic spectrum, thermal imagery, and radar technologies, to accurately measure temporal phenomena at a relatively low cost. It is primarily applied in monitoring crop growth, predicting climate conditions, and overseeing land use changes and water and forest resources. Despite its potential, the majority of these advancements have focused on monitoring legal crops, leaving the increasing problem of illicit crops largely unaddressed.

Notably, in Colombia, approximately 3 % of the cultivated land—about 200,000 of the 7 million hectares—is used for coca cultivation [UNODC and MINJUSTICIA, 2022]. The unique dynamics of this crop demand accurate and scalable modeling. These dynamics include frequent harvest cycles (up to a harvest every 45 days), intensive agrochemical use, regional concentration of cultivation, unique biophysical needs, and high crop mobility due to interdiction [UNODC and MINJUSTICIA, 2010].

Precise, scalable, and cost-effective methodologies for monitoring coca cultivation are crucial for both domestic and international policy-makers. Traditional methods rely on manual visual interpretation of satellite imagery, which is time-consuming, expensive, and susceptible to human error and bias. Therefore, the need for innovative solutions is urgent. To address this, the present study aims to improve the accuracy and scalability of coca crop detection by harnessing the potential of Sentinel-2 imagery and machine learning (ML) [Chen and Guestrin, 2016, Saini and Ghosh, 2021].

The research addresses two central questions: What is the most effective combination of algorithms and features for detecting coca crops using satellite imagery? And, how does the proposed methodology enhance the accuracy, scalability, and continuity of coca crop detection compared to current methodologies? In this context, scalability refers to the capability of any methodology to be applied effectively on a large scale. This includes considerations such as computational efficiency — ensuring that the model can deliver timely predictions without excessive processing time — and the consistent availability and uniformity of required input data across different regions and time periods. By applying this methodology to real-world data from two zones in Putumayo, Colombia during 2019-2021, we aim to assess its practical effectiveness and evaluate its potential for broader application.

This study offers countries, particularly those ill-equipped to monitor and manage coca cultivation, a more affordable and effective tool to comprehend the extent of their problem. It also aims to enhance the capabilities of Colombia's SIMCI (Sistema Integrado de Monitoreo de Cultivos Ilícitos) program, which could significantly contribute to the country's efforts to curb cocaine production. Furthermore, it builds on previous work in the field by examining

the potential of cost-effective imagery and the XGBoost algorithm for large-scale coca crop detection.

The remainder of this document is organized as follows: *Chapter 2* provides a detailed review of the literature on the use of RS and ML in agricultural monitoring, with a particular emphasis on illicit crop detection. *Chapter 3* describes the data and methodology used in this research. *Chapter 4* presents the results of the analysis. *Chapter 5* discusses the findings of the research in the broader context of coca crop monitoring and remote sensing.

This research is poised to make a significant contribution to the field of illicit crop monitoring and the broader efforts to combat the global drug problem. It offers new perspectives and solutions to a complex and persistent issue, situated at the intersection of RS, ML, and policy.

2 Background

2.1. Coca and its agricultural practices in Colombia

2.1.1. Coca species and distribution

The coca plant belongs to the genus Erythroxylum, comprising approximately 200–400 species primarily found in the Neotropic. A minimum of two species have been identified in Colombia: Erythroxylum coca Lam, and Erythroxylum novogranatense, [Galindo and Fernández, 2010].

Ecologically speaking, two primary biophysical and climatological factors determine the optimum conditions for coca cultivation, as described by [Matteucci and Morello, 2001]. In first place, the annual precipitation for the coca crop must be around 2,000 mm, being the range between 1,000 and 4,000 optimal for the crop. As for the altitude, the optimal range is between 1,000 m.a.s.l. and 2,000 m.a.s.l.; although the crop can also thrive below 1,000 m.a.s.l., altitudes 2,000 m.a.s.l are too cold and the photosynthetic activity of the plant stops. When combining these two factors, [Serrano, 2014] shows that almost the entire country has the conditions needed for the crop. This is the main reason why the coca crop is distributed all over the country, affecting each one of the seven regions defined by UNODC.

2.1.2. Cultural and illegal uses

Coca has documented ancestral and cultural uses in the three main producing countries: Colombia, Peru, and Bolivia, linked to indigenous communities. The indigenous practice associated with coca is known as mambeo, which consists of chewing the coca leaves. In Colombia, some indigenous people of the Amazon, the Sierra Nevada de Santa Marta, and the Andean region have declared its use [Ceballos and Lopera, 2009]. Nevertheless, its primary and most problematic use today is for the extraction of an alkaloid known as cocaine. This alkaloid is the main active principle of the plant and was isolated for the first time in 1860. In drug form, cocaine is most commonly snorted as cocaine hydrochloride or smoked as cocaine paste or cocaine base, sometimes referred to as crack.

With the 1961 convention against drug trafficking, and later with the Law 30 of 1986, not

only the coca but other crops that could be used to produce natural drugs were prohibited. The rationale behind these laws and conventions was to address the growing global and regional impact of drug trafficking and its associated social and economic consequences. More regulation was developed after, with the 1988 Vienna convention and its national expression, the Law 67 of 1993. Since then, the coca cultivation is considered an illegal activity with penalties ranging from 5 years and 4 months to 18 years of imprisonment, and with extraordinary fines [Serrano, 2014].

2.1.3. Challenges in coca crop classification

The illegal status of coca cultivation has resulted in its adopting atypical dynamics compared to other agricultural activities, mirroring the patterns seen in other illicit crops. There are three main challenges to be considered in crop classification regarding coca fields:

- High mobility crop. Coca is a highly mobile crop that can rapidly move from one location to another due to eradication and field renewal practices. Although taxonomically coca is a permanent crop, that is, it is a perennial shrub that can live up to 20 years, control activities such as eradication and voluntary or forced substitution programs considerably reduce its permanence time in a fixed place. Along with this, it has been observed that, within the agricultural practices of agricultural producers with coca, it is common to "zoquear" the crop, which implies renewing the crop by cutting the plant from the base [UNODC and MINJUSTICIA, 2010]. This practice serves two purposes. First, it helps maintain high crop yields, which tend to decrease after the fourth year. Second, it's used as a strategy to mitigate the effects of eradication; when the crop is sprayed with herbicide, the producers stifle the crop to prevent its death.

- Agrochemical usage. Another common practice is the intensive use of agrochemicals (fertilizers, herbicides, and pesticides), often with assistance from fellow coca growers, vendors of agricultural input stores, and sometimes even specialized technical assistance provided by illegal armed groups [UNODC and MINJUSTICIA, 2022]. This has allowed the vegetative cycle of the crop to be considerably reduced, achieving more production in less time. In fact, there is evidence that the coca crop can be harvested as often as 45 days, or almost 8 harvests per year [UNODC and MINJUSTICIA, 2010]. Furthermore, thanks to the use of herbicides and pesticides, the level of loss and decrease in production is almost zero, with only the climate and eradication being the main drivers of productivity reduction.

- Mixed cropping. Finally, between 2005 and 2014, the UNODC's yield and productivity studies proved that almost 40 % of coca fields were planted in association with other varieties of coca (25.6 %) or with other licit crops (17.1 %) as informed by

the coca farmers from all the regions in the country. Although this practice is also common in other crops such as coffee, which requires shade provided by species such as the "guamo santafereño"[Farfán and Jaramillo, 2009], for the coca crop, this trend has recently changed since the aerial spraying was suspended in 2015; [UNODC and MINJUSTICIA, 2019] informed that only around 1 % of the coca fields have been associated with other coca varieties or licit crops. This means that in satellite imagery acquired after 2015, the possible challenge of misclassification due to mixed crops will be reduced.

2.1.4. Socioeconomic aspects of coca cultivation

Associated with the cultivation of the crop is the population of coca agricultural producers and their families. [UNODC and MINJUSTICIA, 2010, UNODC and MINJUSTICIA, 2012, UNODC and MINJUSTICIA, 2022] have shown that, at least over the last 15 years, coca producers behave as a mobile population linked to the crop. This means that a large proportion of the agricultural producers with coca come from other municipalities and even other regions than the ones where they are currently located. Research indicates that this group exhibits high levels of illiteracy and low educational achievement. Additionally, the majority of these farmers are male, with the main motivators for cultivating coca being unemployment, poverty, and insecurity.

2.2. The role of remote sensing in coca crop monitoring and classification

Because a multispectral remote sensor captures "the energy reflected or emitted by an object in different bands (regions) of the electromagnetic spectrum", it has been found in the literature that vegetation is no exception, and also falls within the objects, subject to be measured through a sensor. There are bands, or regions of the electromagnetic spectrum, that are sensitive to the spectral response of vegetation. Therefore, in earth image classification, we can use the brightness of individual spectral bands (layers) as independent features, to distinguish different thematic classes. For example, water pixels often correspond to extremely low values in the near-infrared band. According to [Jensen, 2006], in the case of vegetation, its spectral behavior can be measured, indicating the amount of reflective energy in each individual plant along the spectrum. This depends on the nature of the vegetation, its interactions with solar radiation, other climatic factors, and the availability of nutrients and water in its environment. Specifically, in the context of coca crop monitoring, institutions like the UNODC and the US Drug Enforcement Administration (DEA) have used these spectral variations to distinguish coca plants from other types of vegetation and

non-vegetation classes in their respective coca estimation efforts.

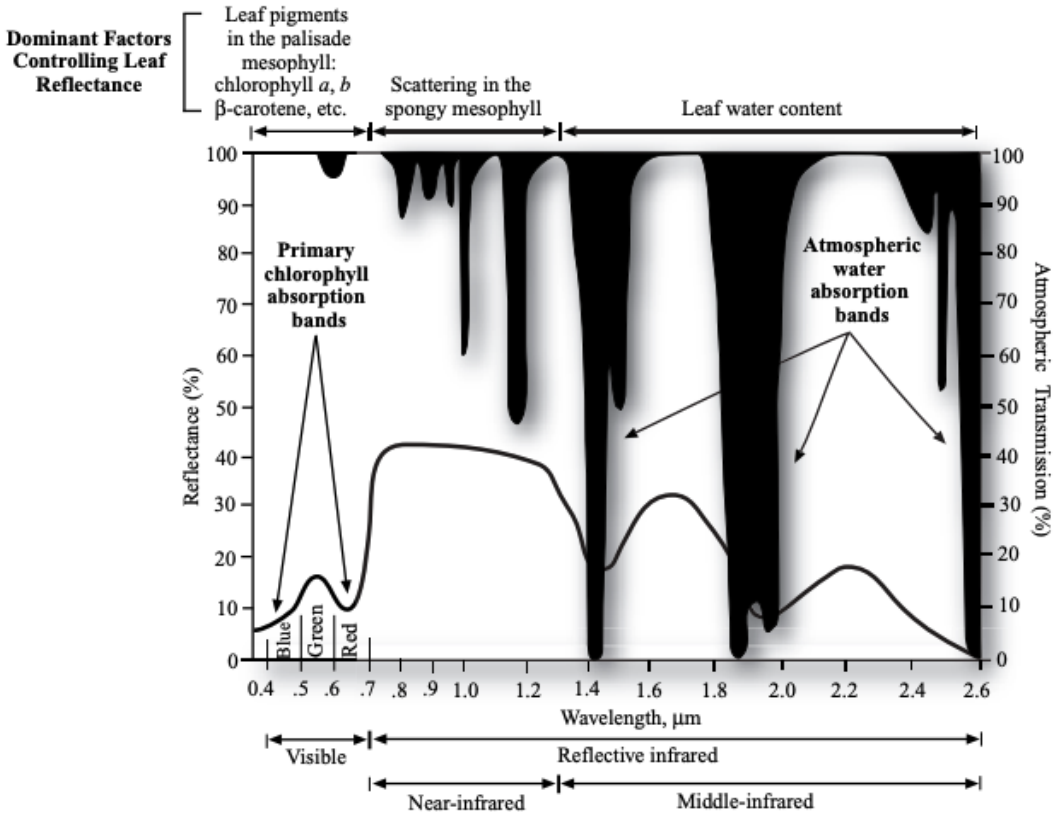

Figure 2-1: Spectral reflectance characteristics of healthy green vegetation for the wavelength range 0.4 - 2.6 μm
Source: [Jensen, 2006]

Classifying land covers and land is one of the most common tasks that can be done using observation imagery [Karpatne et al., 2016]. The most frequent type of classification task is supervised classification. When these methods are used, ground truth class labels also need to be collected to create a training set. This can be done by sending out a field crew on the ground with GPS devices or hiring trained photo interpreters to manually label pixels. Then, a classification algorithm can be used to assign class labels to different image pixels. Extensive research has been conducted on earth imagery classification algorithms. Traditionally, examples of supervised learning methods include maximum likelihood classifiers, decision tree, random forest, support vector machine, and neural network. However, spatial classification and prediction problems pose unique challenges compared to traditional non-spatial classification and prediction problems due to the unique characteristics of spatial data. Authors like [Jiang and Shekhar, 2017] and [Arbia, 2014] propose some of the challenges that can be faced when combining remote sensing with statistical methods and/or machine learning algorithms:

- Spatial autocorrelation (dependency): We cannot assume independence of the observations, since nearby locations can be related to each other. For example, the greenness of a set of pixels that correspond to a same field can be highly correlated. Ignoring spatial autocorrelation can lead to biased estimators and poor prediction performance.
- Spatial heterogeneity: when the study area is broad, the behavior of the same phenomenon can change from one location to another. For example, the spectral, time and spatial feature signatures of coca crops can vary among the different regions of the country.
- Limited ground truth: for supervised classification, label of the units is required. In some cases, this labelling can be outdated or imprecise. Visual interpretation can be both expensive and time-consuming.
- Multiple scales and resolutions: different sensors have different resolutions and scales; combining different sources of information can become a challenge. For example, Landsat imagery has a 30m/pixel resolution while Sentinel-2 has a 10m/pixel resolution.
- Large data volume: every day a lot of information is collected by remote sensors, which makes the data volume to quickly rise. For some analysis done at a large-scale, big-data interfaces and methodologies are needed. For example, SpatialHadoop and Google Earth Engine are platforms that ease the work with multi-petabyte catalog of satellite imagery and geospatial datasets.

It's important to note that in many applications (see [Park et al., 2018], [Tatsumi et al., 2015] and [Piedelobo et al., 2019]), individual spectral bands might not be sufficient to distinguish between different thematic classes. Derived features based on information from multiple spectral bands are often required, as well as time features have been found to be important when the same scene is captured repeatedly over time [Jiang and Shekhar, 2017]. For example, in coca crop monitoring, temporal variations in spectral features could help identify the growth stages of the plants and thus provide more accurate classification results.

Generating derived features has several advantages in earth image classification compared with using spectral band as independent features. First, the derived features can often be interpreted by the physical characteristics of surface spectral response. For example, the vegetation index NDVI can be interpreted by the photosynthetic effect in plants. Second, since derived features are generated without relying on specific training labels, they are often more generalizable than supervised classification models. The reason is that supervised classification methods are typically sensitive to the choice of training samples. The limitation

of generating derived features or indices lies in the difficulties of identifying an appropriate index measure or function. There are multiple spectral bands that may contribute to the phenomena of interest, and these spectral bands may have a nonlinear compound effect. In machine learning, this is called feature engineering. Investigating how to utilize supervised classification methods such as decision trees and deep learning to identify critical derived features or indices for a specific phenomenon is of interest.

2.3. Overview of XGBoost models in remote sensing

In this section, we will provide an overview of XGBoost models, focusing on their application in remote sensing for crop classification. XGBoost is particularly relevant for this application due to its versatility, robustness and ability to handle complex, large scale datasets which are characteristic of remote sensing applications. More specifically, its ability to capture intricate data structures and handle missing values makes it a powerful tool for crop classification tasks, even in challenging conditions. We will first discuss the general methodology of XGBoost, including the relevant equations and a brief introduction to decision trees. Next, we will highlight the advantages of XGBoost, particularly its ability to handle missing values, which is crucial for addressing the presence of clouds in remote sensing images. Finally, we will discuss the most important hyperparameters of XGBoost and present some studies that have applied the method to crop classification.

2.3.1. XGBoost methodology

XGBoost, short for eXtreme Gradient Boosting, is a machine learning algorithm based on the concept of Gradient Boosted Decision Trees (GBDT). In simple terms GBDT is a form of ensemble learning [Friedman, 2001] where a collection of weak learners is trained sequentially to enhance the performance of individual models on specific regression or classification tasks (see **2-2**); it iteratively build a strong model by adding weak learners and fitting them to minimize the objective function, which measures the difference between the predicted values and the true values.

This weak learners are typically shallow decision trees because they usually contain only a few splits. Each of these trees is designed to predict the residuals or errors of the preceding tree. That is, rather than directly predicting the target variable, each tree in the ensemble is trying to correct the mistakes of its predecessor. This approach allows the model to focus on the most challenging parts of the training data, enabling it to incrementally improve its predictions.

XGBoost, proposed in [Chen and Guestrin, 2016], builds upon the concept of GBDT. It introduces computation optimization techniques to the base algorithm. These include a

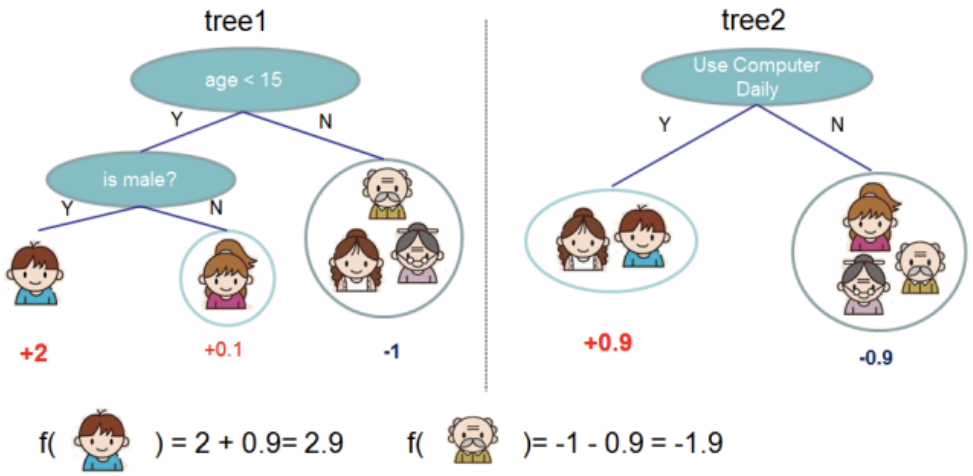

Figure 2-2: Tree Ensemble Model. The final prediction for a given example is the sum of predictions from each tree.
Source: [Chen and Guestrin, 2016]

weighted quantile sketch for efficient proposal calculation, a novel sparsity-aware algorithm for parallel tree learning, and an effective cache-aware block structure for out-of-core tree learning.

For a dataset with n observations and m variables, $D = (y_i, \mathbf{x}_i)|y_i \in \mathbb{R}, \mathbf{x}_i \in \mathbb{R}^m)$, where y_i is the objective of the task, a GBDT model with K learners can be defined as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\mathbf{x}_i), f_k \in \mathcal{F},$$

Where $\mathcal{F} = \{f(\mathbf{x}) = w_{q(x)}\}(q : \mathbb{R}^m \to T, w \in \mathbb{R}^T)$ is the space of regression trees, where q is a possible structure for a tree based with T leaves. Each tree built has an structure q_k, where the T_k leaves have a weight of w_{q_k}.

Each f_k in the ensemble represents a classification and regression tree (CART) [Breiman et al., 1984]. In this context, each $f(x_i)$, where i is the i-th data point in our dataset, represents the output of the k-th classification tree for that data point.

Decision trees, a fundamental component of XGBoost, are a widely used form of machine learning algorithm. They are aptly named for their tree-like structure, in which each node represents a feature (or attribute) and each branch represents a decision rule. These rules lead down to a leaf node which contains the prediction or outcome of the decision tree. By navigating the branches of the tree from the root to a leaf, the algorithm can make a prediction for a given instance based on its feature values.

The original GBDT algorithm built these trees trying to optimize an objective function. In machine learning, the objective function, also known as a loss or cost function, plays a crucial role in training a model. It quantifies how far the model's predictions are from the true values, providing a metric that the algorithm can strive to minimize during the training process. The smaller the value of the objective function, the better the model's predictions. In the context of XGBoost and GBDT, the objective function is defined as follows at each iteration k:

$$\mathcal{L}^k(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{k-1}) + f_k(\mathbf{x}_i) + \Omega(f_k),$$

Where l is a differentiable convex loss function that measures the difference between the prediction $\hat{y}_i$ and the target y_i, and Ω is a complexity penalization that helps the model to avoid overfitting.

XGBoost reformulates this optimization target by using the second order approximation proposed in [Friedman et al., 2000] which allows to speed up the optimization process. With the approximation the new objective function at each iteration k can be considered as:

$$\mathcal{L}^k(\phi) \approx \left[\sum_{i=1}^{n} l(y_i, \hat{y}_i^{k-1}) + g_i f_k(\mathbf{x}_i) + \frac{1}{2} h_i f_k^2(\mathbf{x}_i)\right] + \Omega(f_k),$$

Where g_i and h_i are the first and second order derivates, respectively, of l, respect to $\hat{y}_i^{k-1}$.

Finally, the new optimization function can be translated from the whole model with structure q_k into the space of the leaves, so that the model can choose an optimal leave split for the whole structure based on the membership of each observation to a leaf j, where the set of instances for can be defined as $I_j = \{i | q_k(\mathbf{x}_i) = j\}$). By expanding Ω, we get the resulting optimization function:

$$\mathcal{L}^k(\phi) \approx \left[\sum_{i=1}^{n} l(y_i, \hat{y}_i^{k-1}) + g_i f_k(\mathbf{x}_i) + \frac{1}{2} h_i f_k^2(\mathbf{x}_i)\right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2.$$

After finding the optimal weight for w_j, for a fixed structure q_k, we can create a scoring function that can be used to measure the quality of the structure (see **2-3**), as follows:

$$\tilde{\mathcal{L}}^k(q_k) = \frac{1}{2} \sum_{j=1}^{T} \frac{(\sum_{i \in I_j} g_i)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T.$$

Since ennumeratring all the possible trees structures q_k, a greedy algorithm is implemented, that starts from a single leaf and iteratively adds branches to the tree. Assuming that IL

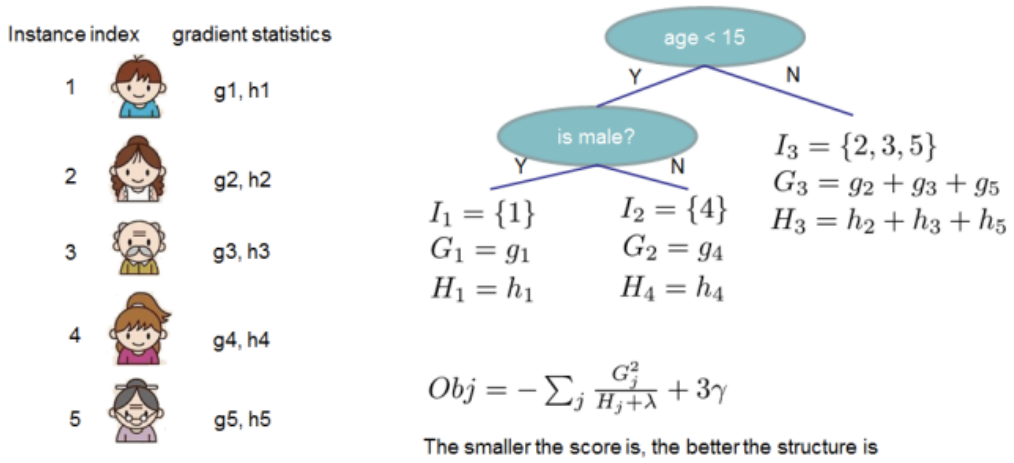

Figure 2-3: Structure Score Calculation. We only need to sum up the gradient and second order gradient statistics on each leaf, then apply the scoring formula to get the quality score

Source: [Chen and Guestrin, 2016]

and IR are the instance sets of left and right nodes after the split, the loss reduction after the split is given by:

$$\mathcal{L}_{split} = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \lambda} + \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma,$$

Which can be used to evaluate the best split candidates.

Based on the improvements in the regularized objective, which have been described, XGBoost introduced several additional techniques to further promote classification performance, with a particular focus on hardware-side optimizations. Notably, the adjustments made to the objective function stand out as the most important addition to XGBoost. These adjustments significantly enhanced the algorithm's ability to handle complex datasets and improve predictive accuracy. Additionally, hardware-side techniques, such as the use of greedy algorithms for split finding, weighted quantile sketch, sparsity-aware split finding, and parallelization, were also introduced to expedite the training process and efficiently utilize computational resources. Readers interested in exploring these techniques in detail can refer to the original source [Chen and Guestrin, 2016]:

- Greedy Algorithms for Split Finding: XGBoost employs a greedy approach to find the best splits during the tree construction process. The algorithm evaluates all possible splits on each feature and selects the one that maximizes the gain in the objective function. This greedy strategy allows XGBoost to efficiently explore the feature space and make informed decisions about split points.

- Weighted Quantile Sketch: The weighted quantile sketch is used to propose possible split points efficiently. It's a technique that approximates the weighted percentiles of a distribution, allowing XGBoost to handle large datasets and make informed split decisions without explicitly examining every data point.
- Sparsity-Aware Split Finding: XGBoost is designed to handle sparse data effectively. Sparse data consists of many zero or missing values, and the algorithm optimizes its data structures and computations to efficiently deal with such situations, leading to faster training times and reduced memory usage.
- Parallelization: XGBoost leverages parallel computing to speed up the training process. It can distribute the computation of tree building across multiple CPU cores or even multiple machines, significantly reducing the overall training time for large datasets.

2.3.2. Advantages of XGBoost

After establishing the general methodology, we will delve deeper into the specific advantages of XGBoost for remote sensing applications, shedding light on its superior performance, scalability, robustness, and ability to handle missing values. XGBoost has several advantages that make it suitable for remote sensing applications:

High accuracy: XGBoost often has been proof to achieve better predictive performance compared to other machine learning algorithms and even to deep learning algorithms, with a fraction of the time and resources employed by other methods (see [Park et al., 2021, Abdikan et al., 2023, Huber et al., 2022]).

Scalability: XGBoost is designed to be efficient and scalable, making it suitable for large datasets commonly encountered in remote sensing (see [Chen and Guestrin, 2016, Huang et al., 2022]).

Handling of missing values: XGBoost can inherently handle missing values by finding optimal splits for the non-missing data and then assigning the missing data to the side of the split that minimizes the loss. In [Chen and Guestrin, 2016] the Sparsity-aware Split Finding algorithm explicitly takings into account the missing values prior to determine the optimal split. This feature is particularly important in remote sensing applications, where the presence of clouds can introduce missing values in the satellite images.

3 Materials and Methods

The study's methodology, utilized for coca crop monitoring using satellite imagery, is succinctly represented in Figure **3-1**. This graphical abstract serves as a visual guide, providing a high-level overview of our machine learning approach. It illustrates the systematic progression of steps followed in both datasets, beginning with the acquisition of satellite imagery, transitioning through stages of feature extraction, model optimization, and culminating in prediction and evaluation of results. This clear representation aids in imparting an intuitive understanding of the workflow followed in this study.

3.1. Data Acquisition and Preprocessing

3.1.1. SIMCI's coca fields and study area

The foundation of this study rests upon a dataset provided by the United Nations Office on Drugs and Crime's Sistema Integrado de Monitoreo de Cultivos Ilícitos (SIMCI). The SIMCI dataset contains detailed information about coca fields interpreted by the PDI team and comprises geographically explicit polygons that denote the spatial extent of coca cultivation areas. Given its comprehensive and accurate representation of coca fields, the SIMCI dataset serves as an invaluable ground truth for monitoring and analyzing patterns of coca cultivation.

The dataset provides information on two distinct zones over a span of three years, resulting in six unique study areas, located within the department of Putumayo, Colombia. The selection of these zones was the outcome of a collaborative effort with the SIMCI team, taking into consideration areas of high-density coca cultivation. These zones provide the geographical context for our two datasets, serving as the primary stage upon which we apply and test our methodologies. For reference, maps of the classes (coca and not coca) of these zones are provided (**3-6**).

In preparation for the analysis, the polygon data from the SIMCI's dataset was transformed into a raster format. We employed QGIS for this rasterization process. The resultant raster maps retain the spatial extent of the coca cultivation areas while enabling seamless integration with the satellite imagery datasets from Sentinel-2 and Planet. This raster data derived from SIMCI's dataset forms the primary target of our work and is used for both

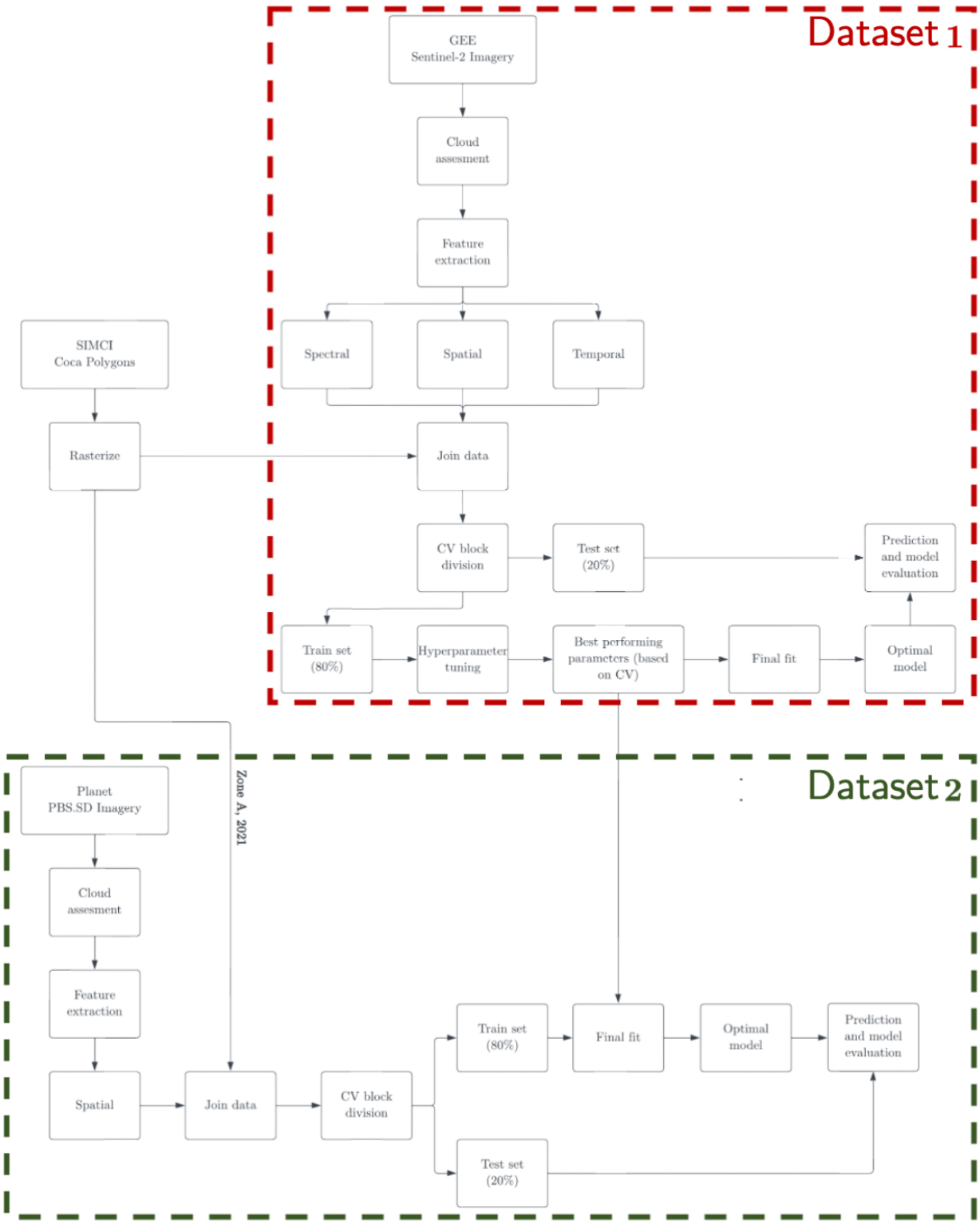

Figure 3-1: Graphical summary of the methodology
Source: Author's own

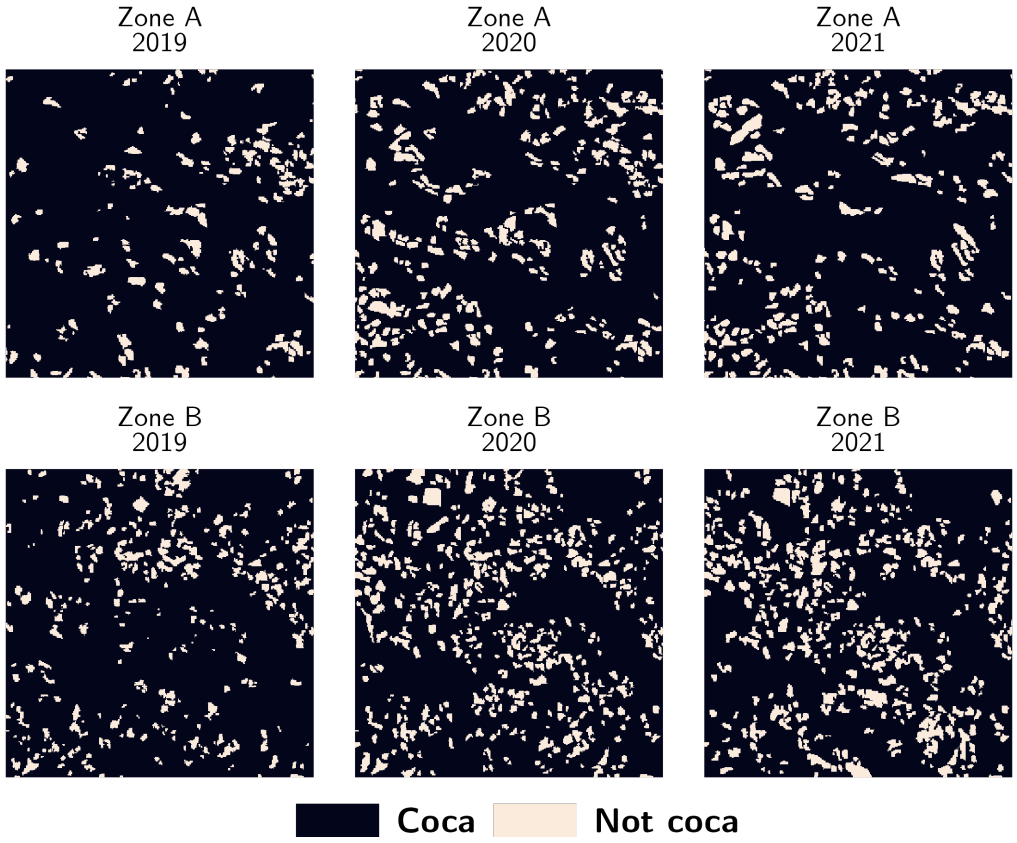

Figure 3-2: Spatial distribution of coca and not coca classes for the study zones across different years

Source: Author's own

dataset.

In order to ensure consistency across the datasets, all our raster data—SIMCI, Planet, and Sentinel—were ensured to use the same projected coordinate system, specifically EPSG:32618, corresponding to UTM Zone 18N.

3.1.2. Satellite Imagery

Each image captured by a satellite is a collection of pixels ordered in a regular squared grid, with each pixel representing a specific area on the ground. Every pixel in these images contains individual values for each spectral band. These values reflect the different characteristics of the surface, such as vegetation, water bodies, soil types, and man-made structures. By analyzing these values, we can derive meaningful insights about the land cover, which is essential for the study on coca crop detection. Specifically, we gather information of two remote sensing developers: Sentinel-2 and Planet.

Sentinel-2

The Sentinel-2 mission, developed by the European Space Agency (ESA), is a key part of the Copernicus Programme and serves as the primary data source for Dataset 1 in this study. Sentinel-2 is equipped with a Multi-Spectral Instrument (MSI) that delivers medium-resolution imagery across 13 spectral bands, spanning from visible and near-infrared (VNIR) to short wave infrared (SWIR)[ESA, nda].
For each study zone and for each year, we downloaded Sentinel-2 imagery spanning an eight-month window. This period starts in June (six months prior to the end of the year) and ends in February (two months following the end of the year). This time frame was selected to align with the reference date of the SIMCI dataset, which specifies coca cultivation status as of December 31st of each year. Given the approximately five-day revisit frequency of Sentinel-2 over Colombia [ESA, ndc], this yielded a dataset consisting of approximately 40 images per zone per year.

Also, we leveraged the power of the Google Earth Engine (GEE) to acquire and process Sentinel-2 data for the selected study zones. GEE, a planetary-scale platform for Earth science data and analysis, provides an efficient way to handle large amount of data for several remote sensing topics like vegetation mapping and monitoring, land cover classification, agricultural applications and disaster managing [Mutanga and Kumar, 2019]. The acquisition and processing was done using `rgee` [Aybar et al., 2020] the GEE extension package built on R. The bands from the Sentinel-2 imagery are detailed in **3-1**.

Name	Pixel Size	Wavelength	Description
B1	60 meters	443.9nm (S2A) / 442.3nm (S2B)	Aerosols
B2	10 meters	496.6nm (S2A) / 492.1nm (S2B)	Blue
B3	10 meters	560nm (S2A) / 559nm (S2B)	Green
B4	10 meters	664.5nm (S2A) / 665nm (S2B)	Red
B5	20 meters	703.9nm (S2A) / 703.8nm (S2B)	Red Edge 1
B6	20 meters	740.2nm (S2A) / 739.1nm (S2B)	Red Edge 2
B7	20 meters	782.5nm (S2A) / 779.7nm (S2B)	Red Edge 3
B8	10 meters	835.1nm (S2A) / 833nm (S2B)	NIR
B8A	20 meters	864.8nm (S2A) / 864nm (S2B)	Red Edge 4
B9	60 meters	945nm (S2A) / 943.2nm (S2B)	Water vapor
B11	20 meters	1613.7nm (S2A) / 1610.4nm (S2B)	SWIR 1
B12	20 meters	2202.4nm (S2A) / 2185.7nm (S2B)	SWIR 2

Table 3-1: ESA's Sentinel-2 band description
Source: [ESA, ndb]

The Sentinel-2 imagery obtained through GEE was already subjected to preprocessing by the data provider, including calibration, atmospheric correction, orthorectification, and co-registration. This ensured that the data was in a ready-to-use state for subsequent analysis.

Planet

The PlanetScope satellite constellation, managed by Planet Labs, provides a significant complement to Sentinel-2 in terms of spatial resolution and serves as the main data source for Dataset 2 in this study. With a pixel size of 3 meters, it provides higher spatial resolution than Sentinel-2, which can be advantageous in studies that require precise localization and detailed analysis at small scales. It captures imagery across eight distinct spectral bands, four of which align with the RGB and Near-Infrared (NIR) bands of the Sentinel-2 instrument. The other four bands —Coastal Blue, Green I, Yellow, and Red Edge— enhance the system's spectral capabilities, with the Red Edge band designed to be interoperable with Sentinel-2 band 5. However, unlike Sentinel-2 data, Planet imagery cannot be processed through the Google Earth Engine (GEE) and it is not freely available [PBC, 2022].The bands from the Sentinel-2 imagery are detailed in **3-2**.

Band Number	**Pixel Size**	**Wavelength**	**Description**
1	3 meters	443nm	Coastal Blue
2	3 meters	490nm	Blue
3	3 meters	531nm	Green I
4	3 meters	565nm	Green
5	3 meters	610nm	Yellow
6	3 meters	665nm	Red
7	3 meters	705nm	Red Edge
8	3 meters	865nm	NIR

Table 3-2: PlanetScope's PSB.SD band description
Source: [PBC, 2022]

To illustrate the importance of image resolution, Figure **3-3** shows how a circle shape is represented when rasterized at different resolutions. At a lower resolution (a), the circle appears more like a cross, whereas at higher resolutions the circle shape becomes more distinguishable (c). This example highlights how resolution can directly impact the perception of shapes in raster data, and by extension, the performance of the model in accurately identifying coca crops.

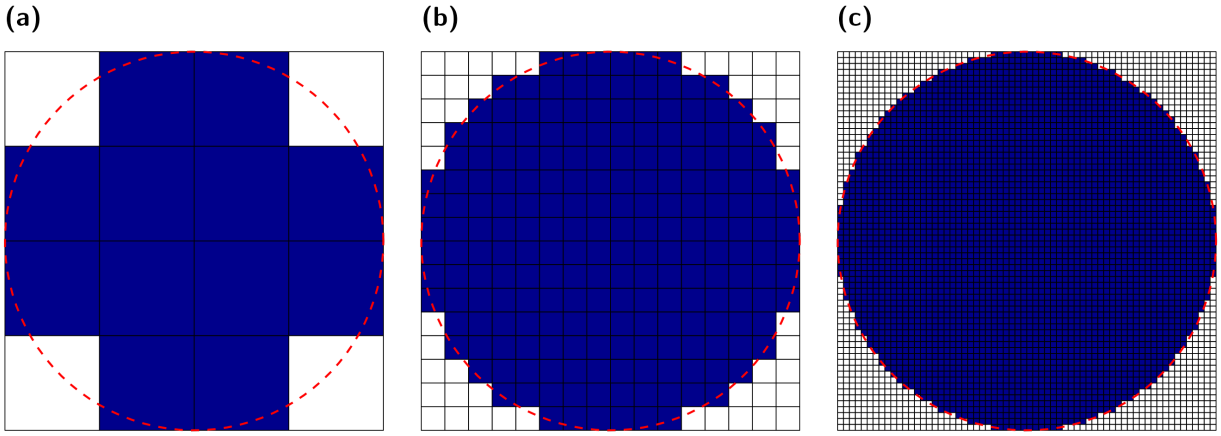

Figure 3-3: Representation of a unit circle in raster data at different resolutions. (a) 1/4u per pixel, (b) 1/16u per pixel, c 1/64u per pixel
Source: Author's own

Figure **3-4** visually demonstrates the difference in resolution between Sentinel-2 and Planet imagery, focusing on the same area of Zone A. The increased level of detail visible in the Planet imagery underlines the potential for improved model performance, as the model can more accurately detect and classify the individual elements present in the high-resolution imagery.

Due to certain restrictions, our study incorporated PlanetScope imagery in a more limited capacity. Specifically, this data was only available for Zone A for the year 2021. The temporal coverage matched that of the Sentinel-2 data, spanning from June 2021 to February 2022. However, due to data availability and quality constraints (i.e., cloud cover), we were only able to obtain one image per month. These images were provided through an agreement with Procalculo Prosis SAS.

As with the Sentinel-2 data, the PlanetScope imagery provided was already preprocessed, including steps such as calibration, atmospheric correction, orthorectification, and co-registration. This preprocessing work by the data provider ensured that the data was immediately ready for our analysis.

3.2. Feature Extraction

In this section, we describe the multiple techniques applied to transform and enrich our original satellite data for subsequent analysis. These techniques fall broadly into the categories of Spectral Indices, Focal Features and Image Aggregation by Month and Temporal

Figure 3-4: Comparison of resolution between Sentinel-2 (left) and Planet (right) imagery
Source: Author's own

Characteristics. Each category has unique benefits and challenges, and their combination serves to create a robust and insightful dataset that provides a comprehensive picture of coca cultivation areas.

3.2.1. Cloud Assessment

The geographical location and diverse climatic and topographic conditions of Colombia make acquiring cloud-free satellite imagery a challenge. For the zones selected in our study, most Sentinel-2 images were largely obscured by clouds. This poses a problem for remote sensing classification as the optical properties of clouds can introduce outlier values, potentially leading to misclassification, such as associating cloud coverage with coca fields.
Our methodology proposes two strategies to address the cloud coverage issue:

- The first strategy uses the GEE platform's cloud segmentation algorithms to identify and exclude cloud pixels from optical sensor images. This method allows

us to avoid the inaccuracies introduced by clouds. However, many models, except for decision tree-based models like Random Forest [Breiman, 2001], XGBoost [Chen and Guestrin, 2016], and LightGBM [Machado et al., 2019], do not support missing values as inputs, limiting our choice of models.

- The second strategy attempts to impute missing values caused by clouds. This requires an evaluation of the impact on accuracy from various methods. We initially focus to compare three imputation methods to complete the observations: kriging for each time [Oliver and Webster, 1990], functional kriging or cokriging [Bohorquez et al., 2017], and an approach proposed by [Gerber et al., 2018] that uses spatio-temporal information and regression models to complete missing values in satellite images.

After considering both strategies, we decided to adopt the first approach, i.e., identifying and assigning missing values to cloud pixels. This decision was due to the time inefficiencies and lack of significant accuracy improvements when attempting to impute nearly empty rasters. It is important to note that while this approach limited our choice of models, the initial intention was to use XGBoost anyway.

3.2.2. Spectral Indices

Spectral Indices are derived from algebraic operations of the different values of the bands for each pixel that enhance the contrast between different types of ground cover and make it easier to distinguish between them [Jensen, 2006]. In the case of Sentinel-2 data, we computed several spectral indices at a pixel-level that are specifically designed to highlight the signature of vegetation, water, and soil in the images, thus making the presence of coca fields more apparent. These indices were added using the `rgeeExtra` package in R, which provides an interface for connecting Google Earth Engine (GEE) and R.

In collaboration with the Procalculo Prosis team, we carefully selected a diverse range of spectral indices that encompass a broad spectrum of phenomena, including vegetation, burning, structures, water, and more. The most commonly used indices, as well as a few others that specifically help in understanding these aspects, were chosen. As a result, the Sentinel-2 imagery was enhanced with the following Spectral Indices:

- **NDCI (Normalized Difference Chlorophyll Index)**: Developed to target chlorophyll content in plant leaves. It is useful in the analysis of crop health. [Mishra and Mishra, 2012]

$$\frac{RedEdge1 - Red}{RedEdge1 + Red}.$$

- **IRECI (Inverted Red-Edge Chlorophyll Index)**: Designed to monitor the chlorophyll content of leaves, which can be indicative of crop health. [Frampton et al., 2013]
$$\frac{RedEdge1 - Red}{\frac{RedEdge1}{RedEdge2}}.$$
- **GVMI (Global Vegetation Moisture Index)**: Provides information on vegetation water content. It can be instrumental in identifying plant stress due to drought. [Ceccato et al., 2002]
$$\frac{(NIR + 0{,}1) - (SWIR2 + 0{,}02)}{(NIR + 0{,}1) + (SWIR2 + 0{,}02)}.$$
- **NDVI (Normalized Difference Vegetation Index)**: A widely used index to quantify vegetation greenness, providing insights into vegetation health and vigor. [Rouse et al., 1974]
$$\frac{NIR - Red}{NIR + Red}.$$
- **SAVI (Soil-Adjusted Vegetation Index)**: Similar to NDVI but includes a soil brightness correction factor, making it more effective in areas with sparse vegetation. [Huete, 1988]
$$\frac{(1{,}428)(NIR - Red)}{NIR + RED + 0{,}428}.$$
- **GLI (Green Leaf Index)**: Designed to highlight the presence of green vegetation. [Louhaichi et al., 2001]
$$\frac{2 \times Green - Red - Blue}{2 \times Green + Red + Blue}.$$
- **NDWI (Normalized Difference Water Index)**: A tool for water detection, used to identify areas of moisture and water bodies. [McFEETERS, 1996]
$$\frac{Green - NIR}{Green + NIR}.$$
- **MTVI2 (Modified Triangular Vegetation Index 2)**: Useful for detecting vegetation cover in areas where the vegetation fraction is low and the soil background has a strong influence on the satellite signal. [Haboudane, 2004]
$$\frac{1{,}5 \times (1{,}2 \times (NIR - Green) - 2{,}5 \times (Red - Green))}{(((2{,}0 \times NIR + 1) \times \times 2) - (6{,}0 \times NIR - 5 \times (Red^{0{,}5})) - 0{,}5)^{0{,}5}}.$$
- **BLFEI (Built-Up Land Features Extraction Index)**: Applied to analyze the reflectance of constructions. [Bouhennache et al., 2018]
$$\frac{((Green + Red + SWIR2) \times 0{,}33) - SWIR1}{((Green + Red + SWIR2) \times 0{,}33) + SWIR1}.$$

While the incorporation of these spectral indices was intended to enhance our satellite data and make it easier to distinguish between coca and not coca pixels, the resulting analysis from our first dataset using Sentinel-2 imagery showed that these variables were not as critical as initially expected. Upon assessing the variable importance, the spectral indices did not significantly contribute to improving the model's predictive performance. This insight led us to omit the computation of spectral indices in the second dataset. Despite this, understanding the role and potential utility of spectral indices in remote sensing and land use classification remains crucial, and future research might explore different sets of spectral indices or alternative methods of application.

3.2.3. Focal Features

In remote sensing, enhancing accuracy often involves integrating spatial characteristics into classification processes. This approach is grounded in Tobler's First Law of Geography, asserting that 'everything is related to everything else, but near things are more related than distant things' [Tobler, 1970]. This law suggests that pixels in proximity, sharing similar characteristics, are likely to belong to the same class.

Focal features capture this concept by considering each pixel as a part of a larger local neighborhood rather than an isolated entity. This neighborhood is delineated by a moving window that surrounds the pixel. Each pixel is then associated with aggregated characteristics of its neighbors, such as the mean, minimum, maximum, and standard deviation of spectral reflectance values. We applied this approach to both basic bands and spectral indices to attain a more detailed understanding of the spatial characteristics present in the imagery.

The moving window's size is an essential parameter to consider: larger windows encompass more distant pixels, while smaller ones focus on the immediate surroundings.

In our study, we implemented the focal features approach for feature extraction, computing the mean, minimum, maximum, and standard deviation of neighboring pixels for each pixel in our datasets. This method has demonstrated efficacy in enhancing accuracy in land cover and land use classification tasks [Chen et al., 2004, Mohanaiah et al., 2013]. It also exhibits significant potential for improving coca field classification accuracy. However, the optimal window size and aggregation function could vary, depending on the specific characteristics of coca fields and the spatial resolution of the imagery, thus offering scope for further research and optimization.

Considering spatial correlations and processing time, we selected a window size of 7 x 7 pixels. This size maintains a balance between capturing local spatial variations and

preserving each pixel's unique attributes. The aim was to incorporate information from both immediate and more distant neighbors, creating a balance between local and broader spatial contexts. This window size is large enough to encompass several coca plants, often clustered together, but also small enough to preserve details. Figure **3-5** illustrates this concept with an example of the mean focal function applied to a 3x3 window.

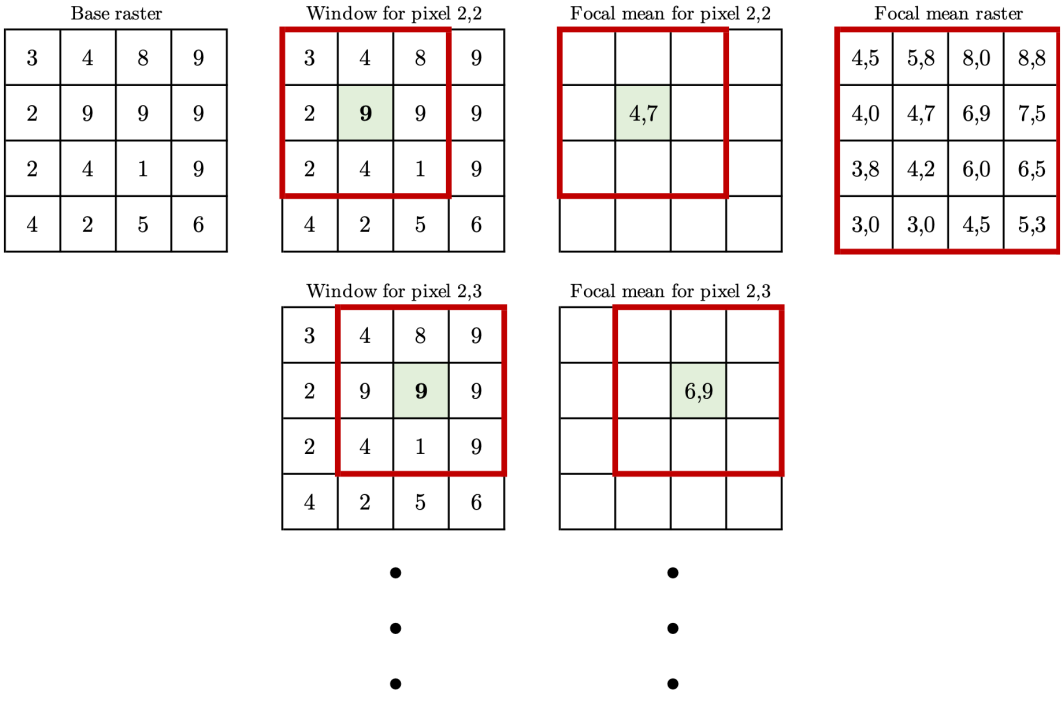

Figure 3-5: Illustration of the mean focal function applied to a 3x3 window
Source: Author's own

3.2.4. Image Aggregation by Month and Temporal Characteristics

Incorporating the temporal aspect of remote sensing imagery is crucial for land cover classification tasks. Plants' phenological cycles, including crops like coca, yield distinct spectral reflectance patterns over time, helping to distinguish them from other types of vegetation [Nazir et al., 2021]. Therefore, considering not only spatial and spectral information but also temporal characteristics of imagery can significantly enhance classification performance [Rustowicz, 2017].

We worked with multi-temporal Sentinel-2 and PlanetScope imagery data, with multiple images available per study area each year. However, managing this data in its raw form can be challenging due to high dimensionality and nearly empty images following cloud removal. Thus, we needed an approach to reduce data dimensionality while preserving essential

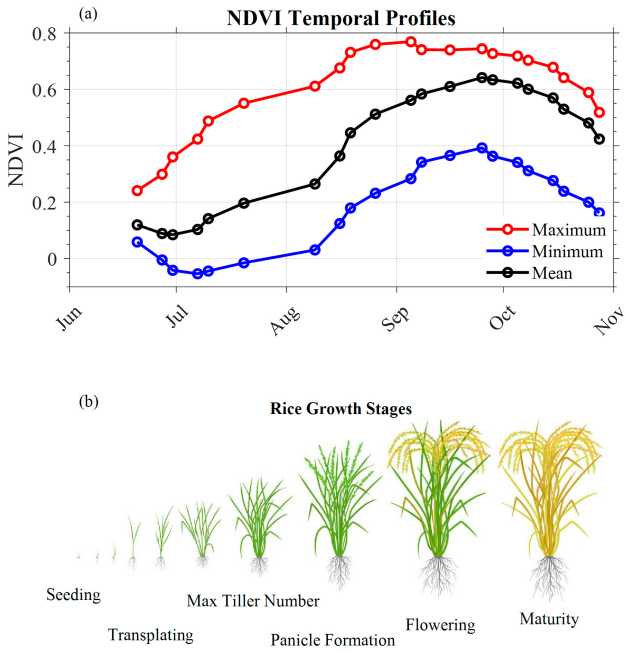

Figure 3-6: (a) Temporal profiles of vegetation index (NDVI) and (b) different growth stages of rice crop
Source: [Nazir et al., 2021]

temporal characteristics.

To address these challenges, we employed the method of creating monthly composites. This technique aggregates data from all images taken within a month into a single image by computing each pixel's average. Repeating this process for all months, we obtained a series of monthly composites representing average conditions for each month of the year. Figure **3-7** illustrates this concept, where three images with missing values due to cloud removal are aggregated.

This approach effectively reduces dataset dimensionality and mitigates the impact of clouds, which can obscure significant image areas. Each pixel in the monthly composites represents multiple observations' average, thus reducing transient phenomena like clouds or atmospheric aerosols' influence.

We created monthly composites for all bands, spectral indices, and their focal versions, ensuring each month's comprehensive land cover characterization. The use of monthly composites preserves the data's temporal dimension, provides insight into the monthly variations in coca fields' spectral properties, and ensures the dataset remains manageable for the subsequent classification task.

Image 1, Month X

3	NA	8	9
2	9	9	9
2	4	1	9
4	2	5	6

Image 2, Month X

8	1	NA	NA
1	2	4	7
1	NA	3	3
7	8	6	1

Image 3, Month X

6	5	4	6
2	NA	8	2
3	3	4	8
1	6	1	NA

Image Composite for Month X

5,7	3,0	6,0	7,5
1,7	5,5	7,0	6,0
2,0	3,5	2,7	6,7
4,0	5,3	4,0	3,5

Figure 3-7: Illustration of month agreggation for a month with 3 images
Source: Author's own

3.3. Hyperparameter Tuning and Cross-validation

XGBoost models, as discussed in the background section, are ensemble learning methods that form a prediction model through an ensemble of weak prediction models, generally decision trees. The model's capabilities are further enhanced by utilizing techniques such as regularization and gradient boosting, which help prevent overfitting and enhance model performance.

The model was implemented using the `xgboost` package in R, providing an efficient and scalable interface to the core C++ code of the XGBoost library, essential for handling the large datasets in remote sensing applications.

XGBoost offers numerous parameters to optimize model performance, but finding an optimal hyperparameter combination depends on the dataset and problem specifics [Yu and Zhu, 2020]. To navigate this complexity, we conducted an exhaustive hyperparameter tuning process using an innovative spatio-temporal approach.

A large grid of hyperparameter combinations was established, involving learning rate, maximum tree depth, subsample ratio, and others. Table **3-3** presents the hyperparameters considered for tuning, their value ranges, and the total unique values each hyperparameter

could assume. The extensive range of hyperparameters and possible values underline the complexity of potential model configurations explored during the hyperparameter tuning process.

Hyperparameter	**Range of Values**	**Number of Unique Values**
Max depth	3:15	13
Eta	0.01 to 0.3, step = 0.01	30
Gamma	0 to 1, step = 0.05	21
Column sample by tree	0.1 to 1, step = 0.05	19
Sub sample	0.1 to 1, step = 0.05	19
Minimum child weight	1 to 20	20
Alpha	0 to 1, step = 0.05	21
Lambda	0 to 1, step = 0.05	21

Table 3-3: Hyperparameters, their ranges of values, and number of unique values in the XGBoost model

Source: Author's own

Considering the vast search space of 26,077,123,800 models, we pragmatically opted to randomly sample 100 hyperparameter combinations for model training. In assessing model performance, standard cross-validation procedures often overlook the spatial-temporal autocorrelation inherent in geospatial datasets. We countered this shortcoming using a novel technique, 'spatiotemporal cross-validation by blocks', inspired by [Valavi et al., 2018]. This method divides the dataset into spatially distinct subsets or 'blocks', providing a realistic representation of the model's spatial predictive capabilities.

This technique serves a dual purpose: it provides an accurate performance metric and demonstrates the model's adaptability across various spatial and temporal scenarios.

In practice, each zone for each year was segmented into five distinct temporal-spatial blocks (Figure **3-8**). All the pixels from the blocks from the first group were reserved for final model evaluation, while the remaining pixels of the four block groups underwent cross-validation during hyperparameter tuning.

In this spatio-temporal cross-validation scheme, the model was trained on three the blocks and evaluated on the fourth in each round. For each one of the 100 hyperparameter combinations that we tried, in the first round, we trained a model using the pixels from blocks 2, 3 and 4 and evaluated on the pixels from block 5. In the second round, we trained a model using the pixels from blocks 3, 4 and 5 and evaluated on pixels from block 2. And so on.

This process was repeated in a round-robin manner until each one of the four blocks had once served as a validation set. The performance metrics for each round were then averaged to yield an overall measure of the model's predictive performance based on the set of hyperparameters used. This strategy ensures the validation data in each round were geographically distinct from the training data, capturing the data inherent spatial autocorrelation, which ultimately leads to a model that can generalize the prediction process on unseen data.

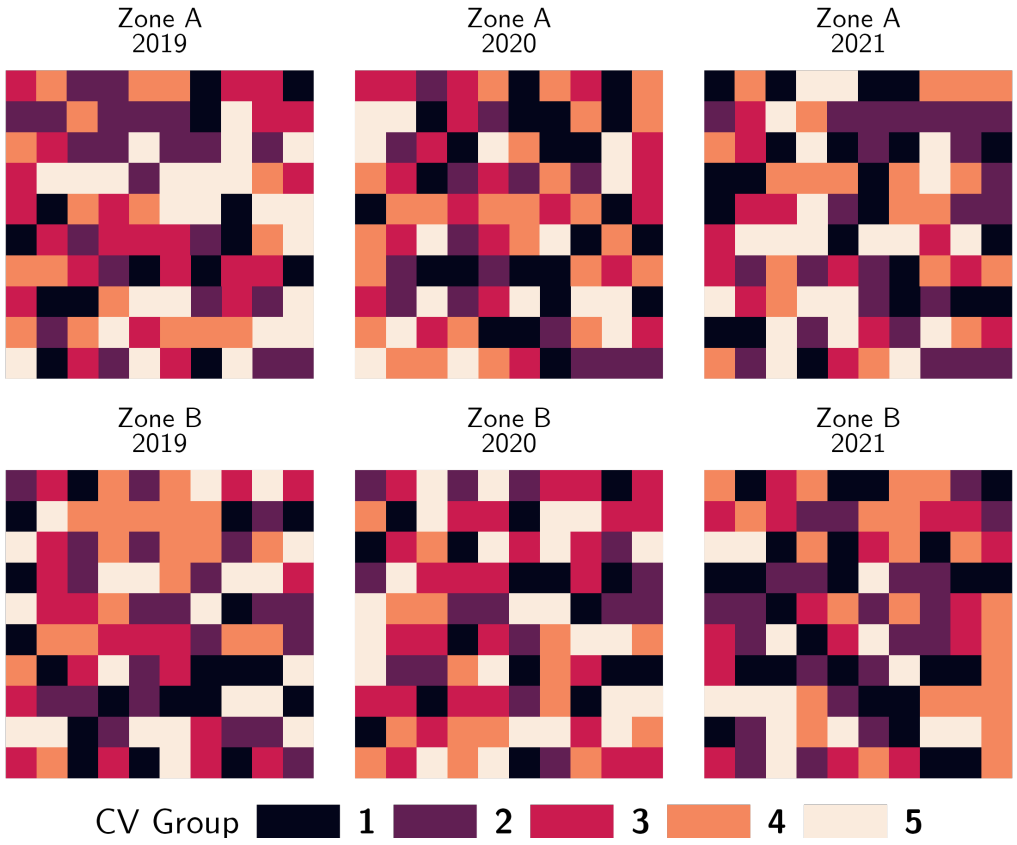

Figure 3-8: Spatial blocks for cross-validation
Source: Author's own

3.4. Tabular Representation of the Data

Although the initial format of the data was raster (.tif) , which is typical in GIS applications, we transformed it into tabular data frames to meet the requirements of the chosen model framework, XGBoost. This format aligns well with traditional data analysis and machine learning techniques, which typically handle data in rows and columns. Using the `terra` we

transformed raster into tables and viceversa.

To adapt our imagery and the extracted features to this format, each pixel from the raster images was converted into a row in our data table. This transformation allowed us to represent spatial, temporal and spectral information from each pixel, including coordinates (X, Y), the label (coca and no coca), the assigned cross-validation group, and the set of extracted features, in a structured and analysis-ready form.

The general scheme for the tabular representation is showned in Table **??**.

Some notes on the tabular representation:

- Each observation (row) is a pixel in a year. Since for each year not only the values of the features changes, but also the label as well, each one of these rows can be considered as an independent observation.
- Each pixel in each year can belong to a different CV group because the blocks of Figure **3-8** are built separately across space and time. However, pixels that are close to each other in space, regardless the time, tend to be in the same CV group. In the schema shown, the 3 unique pixels are consecutive (vertically-wise), so that's why they belong to the same CV group in each year.
- In the base_bands column in the schema, we are representing multiple columns, one for each band and each month. For example, we have the values of the red, green, blue and NIR bands for the months of june, july, etc.
- In the spectral column in the schema, we are representing multiple columns, one for each spectral index and each month. For example, we have the values of the NDVI, MTVI indices for the months of june, july, etc.
- In the focal column in the schema, we are representing multiple columns, one for each band and spectral index, each one of the focal functions and each month. For example, we have the values of the focal maximum (across the neighborhood of the pixel, as described in Section 3.2.3) of the red band in the month of june, or the focal minimum of the NDVI in the month of july, and so on.

3.5. Evaluation Methods

Our classification problem is a highly unbalanced task, wherein the dataset features more examples of the 'not coca' class compared to the 'coca' class in a ratio of about 1:10. This imbalance can result in models that are more sensitive and accurate in classifying the majority class. As our main goal is to build a model that effectively distinguishes between

'coca' and 'not coca', we adopted a suite of metrics to evaluate the models' performance, with a specific focus on penalizing the imbalanced classification. These metrics include Overall Accuracy (OA), F1 Score, and the Kappa statistic:

$$Overall\ Accuracy\ (OA) = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F\ Score\ (F1) = \frac{2 \times P \times S}{P + S}$$

$$Kappa\ (\kappa) = \frac{TP \times TN - FP \times FN}{(TP + FP) \times (FP + TN) + (TP + FN) \times (TN + FN)}$$

Where:

- TP are true positives, pixels that were coca and were correctly classified as coca
- TN are true negatives, pixels that were no coca and were correctly classified as no coca
- FP are false positives, pixels that were no coca and were incorrectly classified as coca
- FN area false negatives, pixels that were coca and were incorrectly classified as no coca

These metrics quantify the models' effectiveness in distinguishing between the 'coca' and 'not coca' classes, especially given the high class imbalance in our dataset.

Aditionally, when we used lanet Labs imagery, an additional evaluation metric was introduced to assess the spectral homogeneity of classified areas. This assessment was based on the computation of the Van Valen's Multivariate Coefficient of Variation (MCV) for both reference and model-classified polygons that overlapping (see [Valen, 1974, Aerts et al., 2015]). The Van Valen's MCV can be expressed as:

$$\gamma_{VV} = \sqrt{\frac{tr\Sigma}{\mu^t \mu}}.$$

A lower MCV value means greater spectral homogeneity, indicating that the pixels within the corresponding polygon show more consistency in their spectral characteristics. By comparing the MCVs of reference polygons with those of model-classified polygons, we evaluated whether the model was generating more spectrally consistent classifications.

While the primary metrics (Overall Accuracy, F1 Score, and Kappa) focused on pixel-level accuracy, the MCV measure provided insight into the model's ability to generate spectrally

consistent classifications at a broader, polygon level. This comprehensive evaluation strategy offered a holistic understanding of model performance, thereby strengthening the robustness of our coca classification methodology.

4 Results

4.1. Dataset 1: Sentinel-2 imagery

4.1.1. Sentinel-2 setup

We utilized the Sentinel-2 imagery including the extracted features as described in Subsections 3.2.2 and 3.2.3. With this data, and using the spatio-temporal block cross validation methodology we identified an optimal set of hyperparameters. With this specific hyperparameters the model achieved the lowest average log-loss function values during the cross-validation process among the 100 possible combination that were tested, indicating the best balance between model complexity and prediction accuracy. The optimal hyperparameters were:

- Max depth = 14
- Eta = 0.03
- Gamma = 1
- Column sample by tree = 0.65
- Subsample = 0.9
- Minimum child weight = 10
- Alpha = 0.4
- Lambda = 1

4.1.2. Model performance

The model, with these hyperparameters, performed exceptionally well on the training data, with an Overall Accuracy (OA) of 99.6 % and an F1 score of 99.8. This means that the model was almost perfect in identifying and classifying the training data. High accuracy and F1 score on the training set demonstrate the model's ability to learn from the data and effectively capture the relationship between the features and the target variable.

In terms of the model's performance on the test data, the Overall Accuracy was 94.3 %, and the F1 score was 76.7. While there was a slight decrease in performance compared to the training data, these scores are still quite high. The slight drop in performance between training and test datasets is a common occurrence in machine learning models and is usually attributed to the model's ability to generalize to unseen data.

When compared to previous work in the field, these performance metrics show and improvement. The most similar study to ours was conducted by [Ángel, 2012], who utilized high-resolution imagery to classify coca crops and achieved a Kappa score of 0.709. In contrast, our model outperformed this previous work by achieving a Kappa score of 0.7512 on the test set, highlighting the efficiency and cost-effectiveness of our methodology. This success demonstrates the potential of our approach to provide an accurate and affordable means for coca crop identification, using freely available Sentinel-2 imagery.

Prediction	Reference	
	Not Coca	Coca
Not Coca	1,038,567	133
Coca	4,748	155,372

Table 4-1: Confusion matrix of the training data for Dataset 1

The confusion matrix of the training data indicates that the model correctly predicted 1,038,567 not coca pixels and 155,372 coca pixels. However, there were 4,748 instances where the model incorrectly classified not coca pixels as coca (false positives), and 133 instances where the model incorrectly classified coca pixels as not coca (false negatives). This demonstrates the model's ability to capture the complex patterns of coca cultivation in the training data, although there were still minor misclassifications.

Prediction	Reference	
	Not Coca	Coca
Not Coca	251,219	8,987
Coca	8,212	31,262

Table 4-2: Confusion matrix of the test data for Dataset 1.

The confusion matrix of the test data shows that the model correctly predicted 251,219 not coca pixels and 31,262 coca pixels. However, there were 8,212 false positives and 8,987 false negatives. While the number of misclassifications increased in the test data compared to the training data, the model's performance remained commendably high. Furthermore, despite the model not having 100 % accuracy, there is a balance between the types of

misclassifications. This ensures that the total coca reported is reasonably close to reality, a desirable property of the model.

4.1.3. Mapping

To provide a spatial representation of the XGBoost models coca crop predictions based on Sentinel-2 imagery, we created two types of maps: a probability map and a confusion map. In the creation of the confusion maps, a default threshold of 0.5 was applied to classify pixels as coca or not coca.

The probability map (Figure **4-1**) displays the probability of coca cultivation at each pixel, as predicted by the model. Brighter areas correspond to higher probabilities of coca cultivation, and these bright zones align closely with known coca cultivation areas. Lower probability areas, which appear darker on the map, coincide predominantly with not coca zones such as forest areas. This spatial distribution supports the model's accuracy in discriminating coca crops from not coca areas.

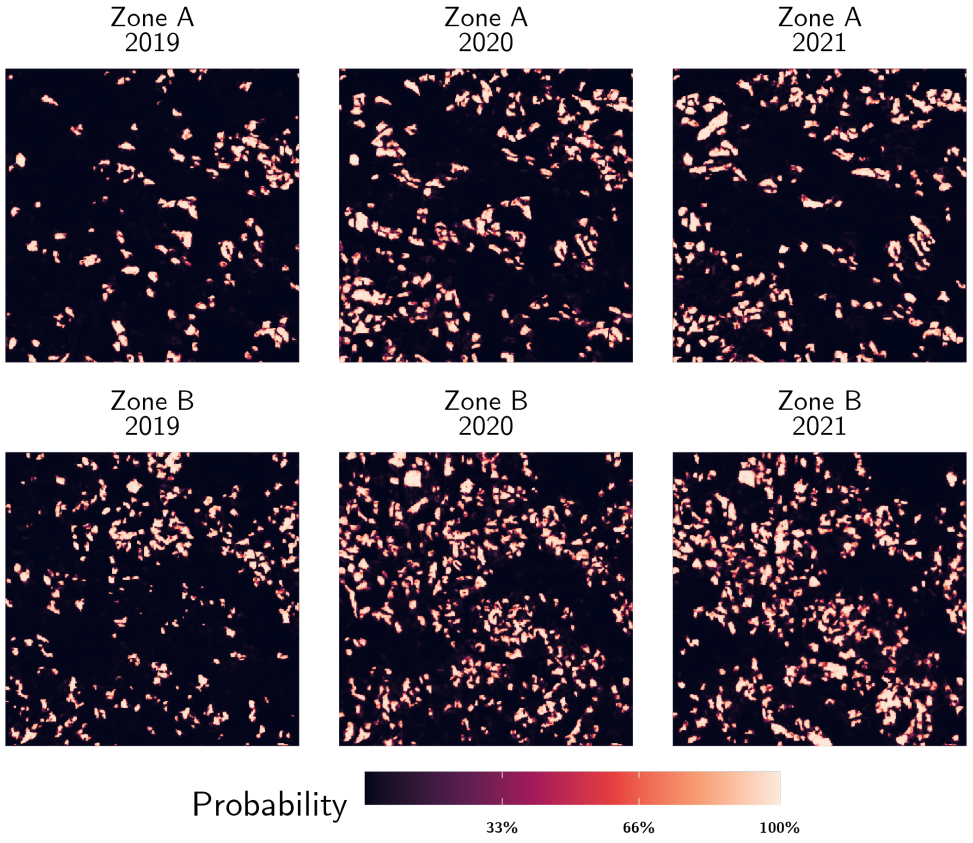

Figure 4-1: Probability map of coca for Dataset 1
Source: Author's own

The confusion map (Figure **4-2**) provides a more granular understanding of the model's

predictions. It highlights the locations of true positives, false positives, true negatives, and false negatives. The map showcases the model's prowess in accurately identifying both coca and not coca areas, with the majority of misclassifications occurring in regions where coca and not coca areas overlap, primarily at the boundaries of the polygons. These boundary regions often present complex and mixed spectral signatures that can challenge the model's ability to correctly classify the pixels.

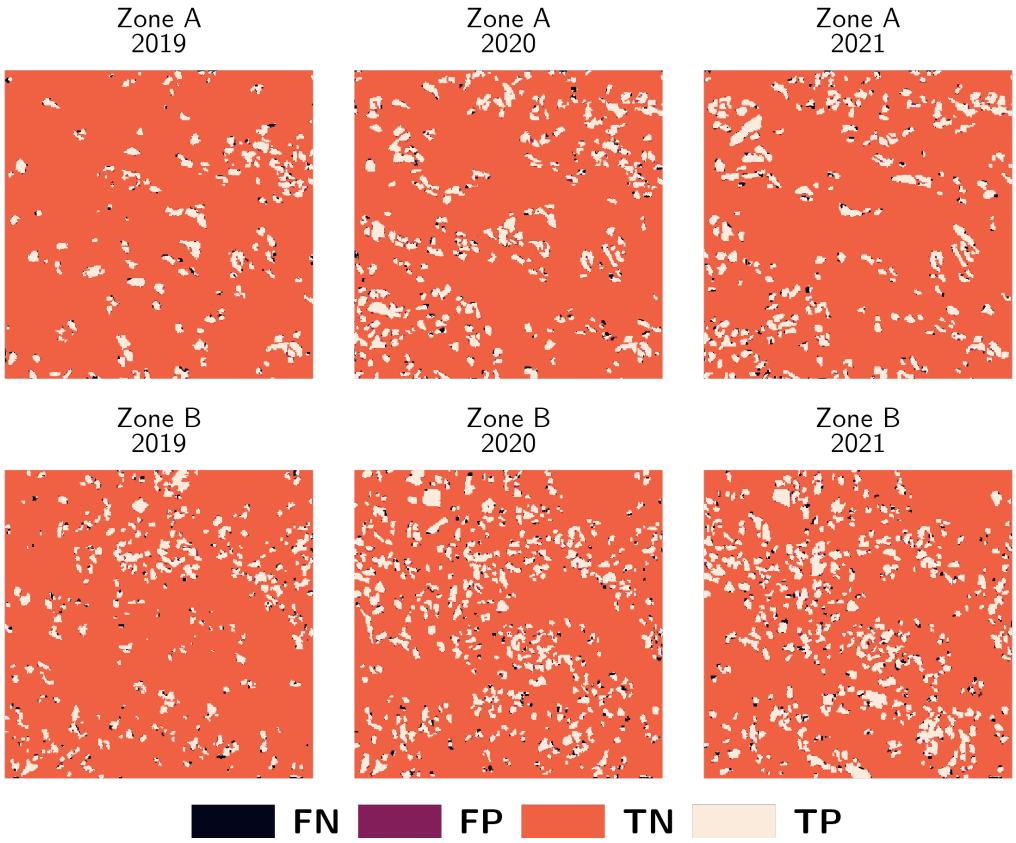

Figure 4-2: Confusion maps for Dataset 1
False Negatives (FN); False Positives (FP); True Negatives (TN); True Positives (TP)

Source: Author's own

The close resemblance of these maps to reality emphasizes the model's high performance and its potential for practical application in coca crop detection and monitoring. This visualization aids in the understanding of the spatial distribution of coca cultivation and model's prediction errors, an understanding that can inform future improvements in the model and refinement of the prediction process.

4.1.4. Variable importance

An essential aspect of understanding the XGBoost model's performance is examining feature importance. In this context, feature importance is assessed based on the average Gain, which is the average improvement in accuracy brought by a feature to the model's predictions. Specifically, Gain is calculated as the difference in loss function value before and after a split based on a feature. A higher average Gain value indicates a more substantial impact of the feature on the model's predictive accuracy, reflecting its relevance in the construction of the decision trees within the model. The overall importance of a feature is then determined by averaging the Gain values across all the trees where the feature is used.

The variable importance of the Sentinel-2 XGBoost model, evaluated by the Gain, is shown in Figure **4-3**. The 'oct_B12_mean_max' feature exhibits the highest gain, underscoring its pivotal role in the model's predictive accuracy. This feature is the focal maximum of the B12 band for the month of October. Its high Gain implies a substantial enhancement in model accuracy wherever this feature contributes to decision-making, thus marking its importance in the model's predictive capability.

Notably, several of the top features are related to band 12 (e.g., 'oct_B12_mean_max', 'aug_B12_mean_max', 'sep_B12_mean_max', 'jan_B12_mean_max'), a shortwave infrared band in the Sentinel-2 imagery. This band captures essential information about vegetation and water content, which appears to be crucial for the classification of coca crops, even though this observation has not been reported in any of the previous studies that we consulted.

The 'x' and 'y' features, which represent the spatial coordinates of the pixels, also rank among the top influential features according to the gain values (see Figure **4-3**). This highlights the significant role of geospatial context in our model's predictions. The model appears to have discerned spatial relationships or patterns that are crucial for coca cultivation. Unraveling these spatial patterns could be a promising direction for future research and might potentially enhance the model's predictive power.

Contrary to expectations, none of the spectral indices typically employed in remote sensing studies feature among the most important variables. This suggests that the XGBoost model, due to its inherent capability to capture complex, non-linear relationships between variables, may effectively extract and use raw spectral information without the need for these preprocessed indices. This finding lends support to the assertion that machine learning techniques such as XGBoost might be better equipped to utilise raw spectral information from remote sensing data, as compared to traditional classification approaches that rely heavily on spectral indices.

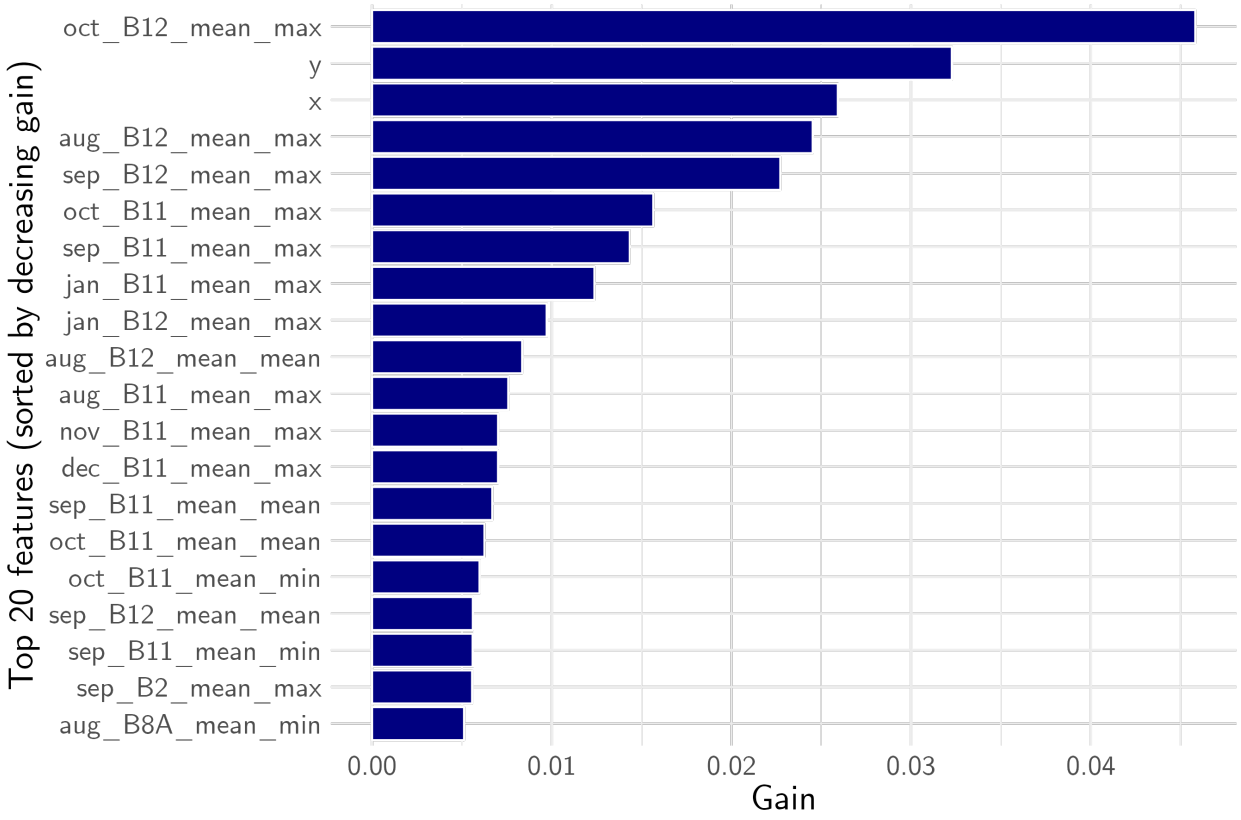

Figure 4-3: Variable importance of top 20 features for the Dataset 1 best model, evaluated by the Gain

Source: Author's own

4.2. Dataset 2: Planet imagery

4.2.1. Planet setup

We used high-resolution Planet imagery to further test the robustness and validity of our model. The motivation behind this choice was two-fold:

First, we aimed to address potential limitations associated with the medium-resolution of Sentinel-2 imagery, which could have influenced the performance of the model. We hypothesized that the medium resolution might have inadvertently caused our model to rely excessively on the SIMCI's data. Since SIMCI's data is generated through image interpretation, it carries a level of subjectivity which may impact the model's ability to generalize.

Second, by introducing high-resolution Planet imagery and implementing more precise geometric delineation (by excluding border areas), we sought to enhance the model's capability to identify coca crops more accurately. This step was designed to ensure that our model was learning and generalizing, rather than merely echoing patterns from the training data.

Due to data limitations, we focused solely on Zone A for the year 2021. Figure **4-4** presents the spatial distribution of the 'coca,' 'not coca,' and 'excluded' classes used for Dataset 2. This figure highlights the specific areas within Zone A that were included in the analysis, and those intentionally excluded to achieve a more precise geometric delineation of coca crops.

It is also crucial to note that for this dataset, we did not perform any further hyperparameter tuning. Instead, we utilized the best-performing hyperparameters identified in Dataset 1. Moreover, in light of the results from Dataset 1, we decided to exclude the spectral indices as they did not appear to contribute significantly to the model's performance.

4.2.2. Model performance

The results of Dataset 2, as outlined in Table **4-3**, indicate a considerable improvement in the model's performance when both temporal and spatial features are included. The performance metrics, Overall Accuracy (OA) and F1 score (which combines precision and recall), were utilized for model evaluation.

When we used only the base features from December 2021, the model attained an OA of 84.6 % and an F1 score of 44.5 on the training set. The test set demonstrated an OA of 83.6 % and an F1 score of 38.8.

The addition of temporal features resulted in substantial enhancement in the model's performance. The OA on the training set climbed to 96.3 %, and the F1 score reached 85.8. Concurrently, for the test set, the OA improved to 91.9 %, and the F1 score rose to 67.0.

The incorporation of spatial features led to a modest increase in the model's performance. The training set observed an OA of 86.3 % and an F1 score of 50.5. The test set performance was slightly better, with an OA of 84.8 % and an F1 score of 42.6.
However, the most pronounced increase in model performance was evident when both temporal and spatial features were integrated. The model exhibited an impressive OA of 99.2 % and an F1 score of 97.2 on the training set. Similarly, on the test set, the model demonstrated an OA of 93.5 % and an F1 score of 75.6. These results point to superior

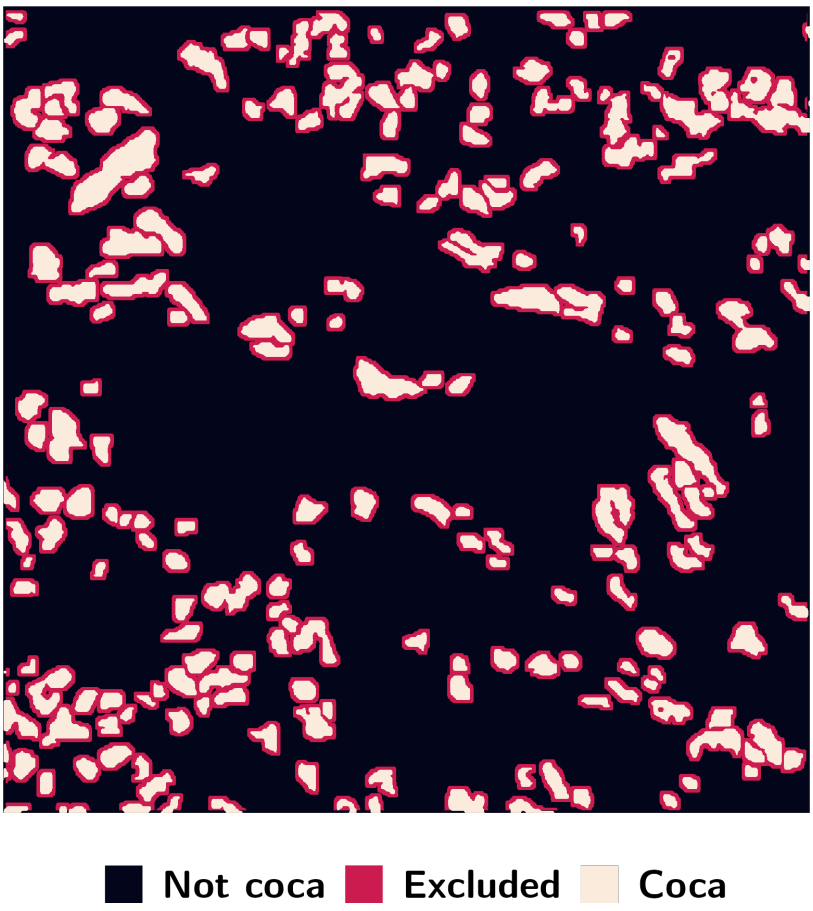

Figure 4-4: Spatial distribution of coca, not coca and excluded classes for Dataset 2
Source: Author's own

model performance and generalization capabilities when both temporal and spatial features are incorporated.

These findings highlight the crucial role of integrating both temporal and spatial information in the model when dealing with high-resolution Planet imagery. The amalgamation of these feature types appears to enable the model to capture more complex and insightful patterns, thereby leading to enhanced precision and reliability in detecting coca crops.

4.2.3. Mapping

To further illustrate our model's performance, we generated two maps using the best model: one depicting predicted probabilities of coca cultivation, and another showing the confusion matrix of the predictions. These visualizations allowed us to assess the similarity between the model's predictions and the reference data used for training.

Dataset	Metric	Base (Dec. 2021)	Time	Space	Time + Space
Training	OA	84.6	96.3	86.3	**99.2**
	F1	44.5	85.8	50.5	**97.2**
Test	OA	83.6	91.9	84.8	**93.5**
	F1	38.8	67.0	42.6	**75.6**

Table 4-3: Model performance by features included (Planet)

Figure **4-5** displays the probability map of coca cultivation based on our model's predictions. Brighter regions correspond to areas where the model predicts a higher likelihood of coca cultivation. These regions align closely with the known distribution of coca cultivation in this area.

In contrast, areas depicted in darker shades, which indicate lower probabilities of coca cultivation, largely coincide with regions known not to support coca cultivation, especially forest zones. This concordance suggests that the model effectively discriminates between forests and coca crop areas.

Regions with intermediate probabilities suggest a degree of uncertainty in the model's predictions. These areas often correspond to geographic features visually similar to coca fields, such as other types of crop fields and field borders.

Figure **4-6** provides a spatial representation of the confusion matrix of the predictions, demonstrating True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). This visualization offers further insight into the model's areas of success and challenge. Most errors occur at the borders of coca fields, which indicates a high degree of accuracy in the model's ability to recognize coca crops within the fields themselves.

Interestingly, confusion at the field borders might not necessarily point to a limitation of the model. On the contrary, it might highlight an issue with the geometric precision of SIMCI's polygon data. Our model was trained to identify coca crops and not specifically border regions. Hence, discrepancies at the borders might underscore the precision mismatch between the high-resolution Planet imagery and the less precise SIMCI polygons. This reinforces our initial hypothesis that the model, when trained with medium-resolution Sentinel-2 imagery, could have over-relied on the SIMCI interpretation, which might be deficient in geometric precision.

For the second dataset, our approach was deliberately more exploratory. We wanted the model to be less complacent and more independent, to truly 'learn' and accurately identify coca crops rather than simply echo the reference data. Hence, we were not deterred by the

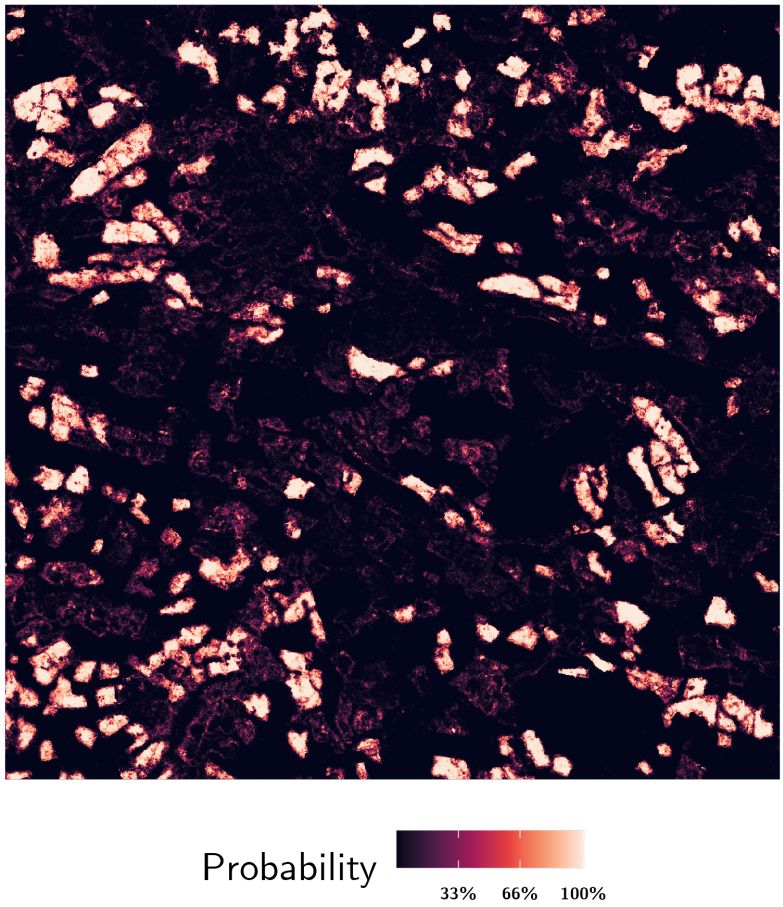

Figure 4-5: Probability map of coca for Dataset 2
False Negatives (FN); False Positives (FP); True Negatives (TN); True Positives (TP)
Source: Author's own

lower accuracy achieved with the high-resolution Planet imagery. On the contrary, we viewed this as a sign of the model learning to discern complex patterns and generalizations, instead of overfitting to the training data. The improved performance with high-resolution Planet imagery underscores this point, suggesting that the model is indeed capable of learning to identify coca crops with a high level of accuracy and independence. This exploratory approach, embracing uncertainty and complexity, was the basis to further explore the hypothesis on geometric precision improvements delivered by this second approach.

4.2.4. Geometric precision improvement

To quantitatively evaluate the model's ability to delineate coca crops with superior geometric precision compared to the SIMCI data, we employed an analysis using Van

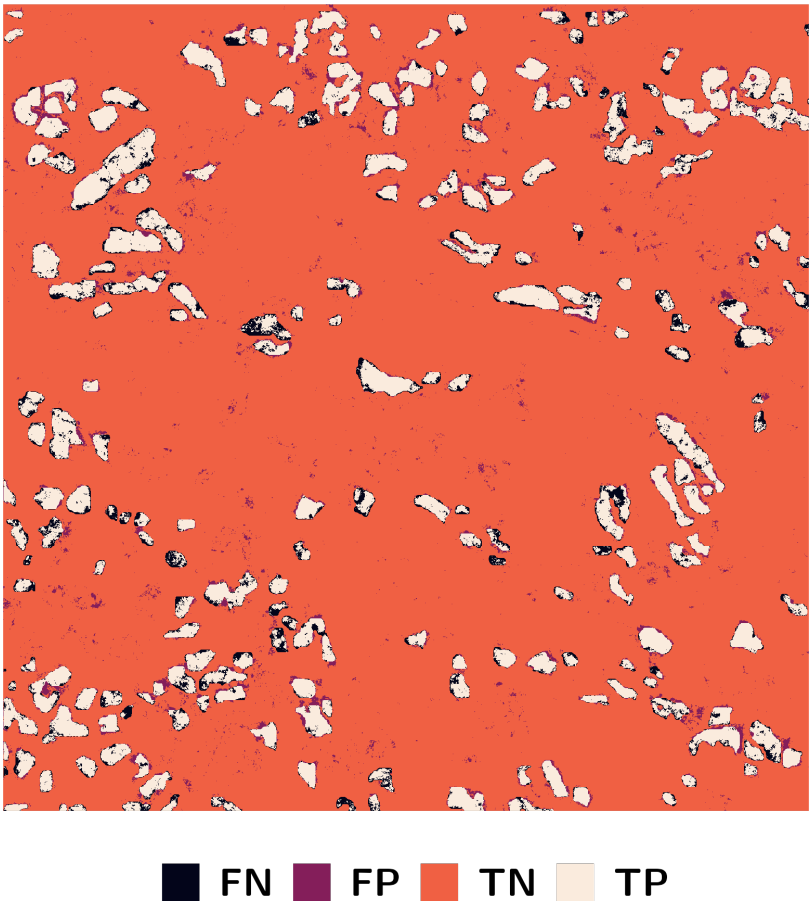

Figure 4-6: Confusion maps for Dataset 2
Source: Author's own

Valen's Multivariate Coefficient of Variation (MCV). The SIMCI data included 217 reference polygons, while the XGBoost model identified 198. This discrepancy suggests that there isn't a one-to-one correspondence between the SIMCI's and the model's classifications, which may be attributed to two potential reasons: envelopment and omission.

Envelopment suggests that one of the model's identified polygons might overlap with more than one of the reference polygons from SIMCI's data. Essentially, this indicates that the model might be combining multiple coca crops into a single prediction.

On the other hand, omission implies that the model failed to identify a polygon that was present in the reference data. This could occur due to the model missing a coca crop field or misclassifying it. These omissions could be seen as a byproduct of the model's advanced learning from the high-resolution Planet imagery, enabling it to discern coca crops with higher accuracy and independence. These discrepancies between the SIMCI and the model's

predictions serve as the foundation for our future work to improve the model's geometric precision in identifying coca crops.

Figure **4-7** illustrates a clear example of the model's superior delineation of a polygon, compared to the SIMCI data. The boundaries defined by the model align more closely with the actual geographic features of the coca crop, demonstrating its superior geometric precision.

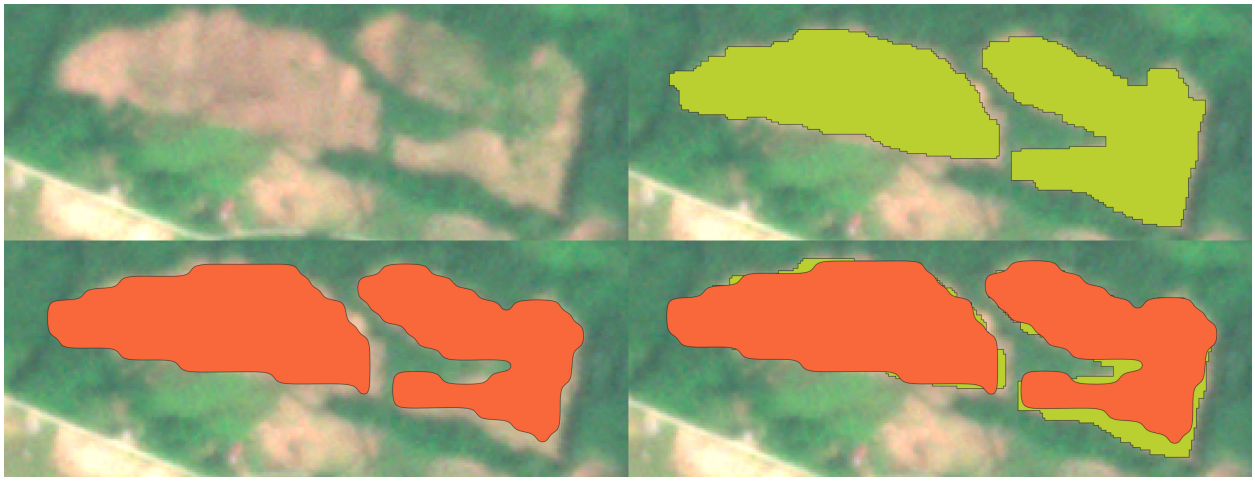

Figure 4-7: Example of improved delimitation. Greens polygon are classified by the model. Red polygons are the SIMCI's reference
Source: Author's own

In contrast, Figure **4-8** shows an example where the model's delineation falls short. The XGBoost polygon in this case encompasses multiple SIMCI polygons, resulting in a loss of precision.

Out of the 212 overlapping polygons between the SIMCI data and the model's predictions, the model was able to reduce the variability in 180 cases (84.9 %). In contrast, the MCV increased in 32 cases (15.1 %), suggesting that the model's predictions don't always perfectly align with the SIMCI data. On average, we observed a decrease in the MCV by 16.39 %. Particularly, the model struggled with cases where multiple SIMCI polygons were encompassed by a single model-generated polygon. In three instances, we found that our model had encompassed four SIMCI polygons, similar to the case illustrated in Figure **4-8**.. These instances resulted in the most substantial loss of precision, with an average increase in the MCV by 23.84 %.

Figure **4-9** presents a comparison of MCV values for SIMCI polygons and their corresponding model polygons.

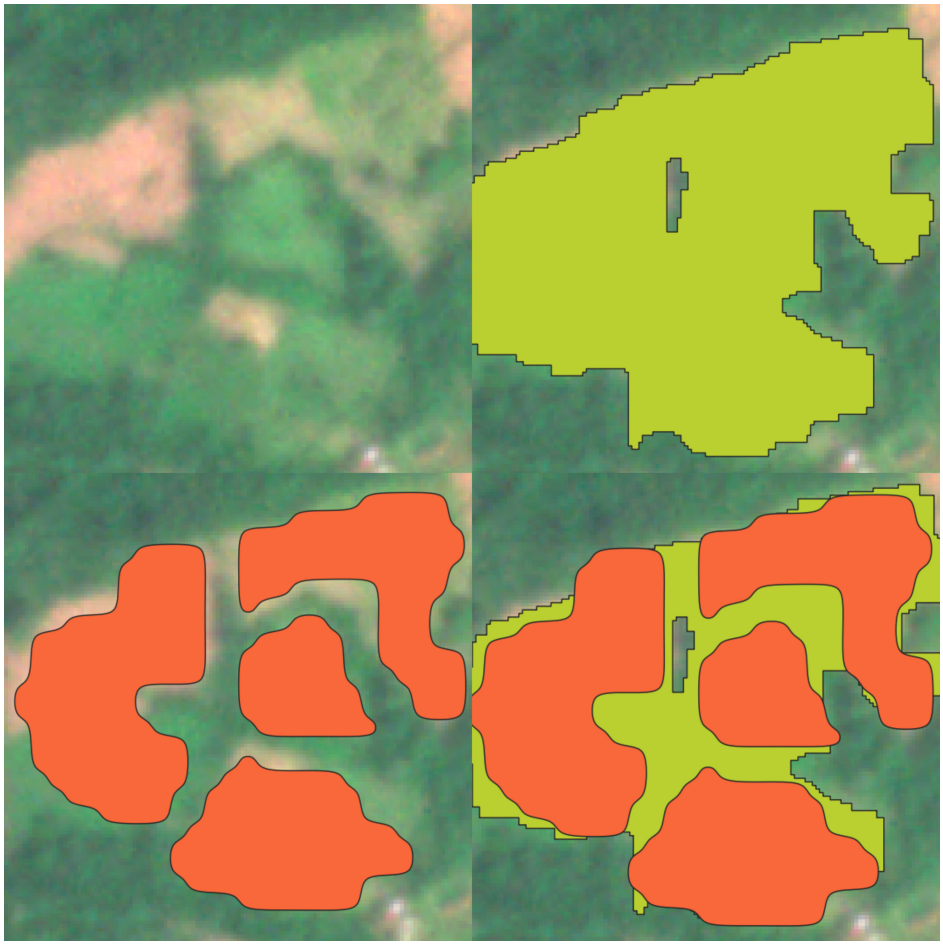

Figure 4-8: Example of worse delimitation due to envelopment. Green polygons are classified by the model. Red polygons are the SIMCI's reference
Source: Author's own

The observed synchronization of growth and harvest cycles across all polygons within a given area was surprising. It was initially believed that each polygon would have its own independent growth and harvest cycle. However, the observed patterns as depicted in Figure **4-10**, indicate that there may be a level of synchronization among different coca crops within the same geographical area. This pattern of increase during the months 8 and 11, followed by a subsequent fall, aligns well with the expected crop growth and harvest cycles, offering a potential temporal feature for the identification of coca crops.

Interestingly, no significant differences were observed between the temporal profiles of the SIMCI and XGBoost groups, further bolstering the validity of our model. This not only reinforces our model's capability to accurately identify and delineate coca crops but also underlines the importance of considering temporal features when developing and deploying models for crop identification.

These results serve as strong confirmation of our model's capability to accurately identify

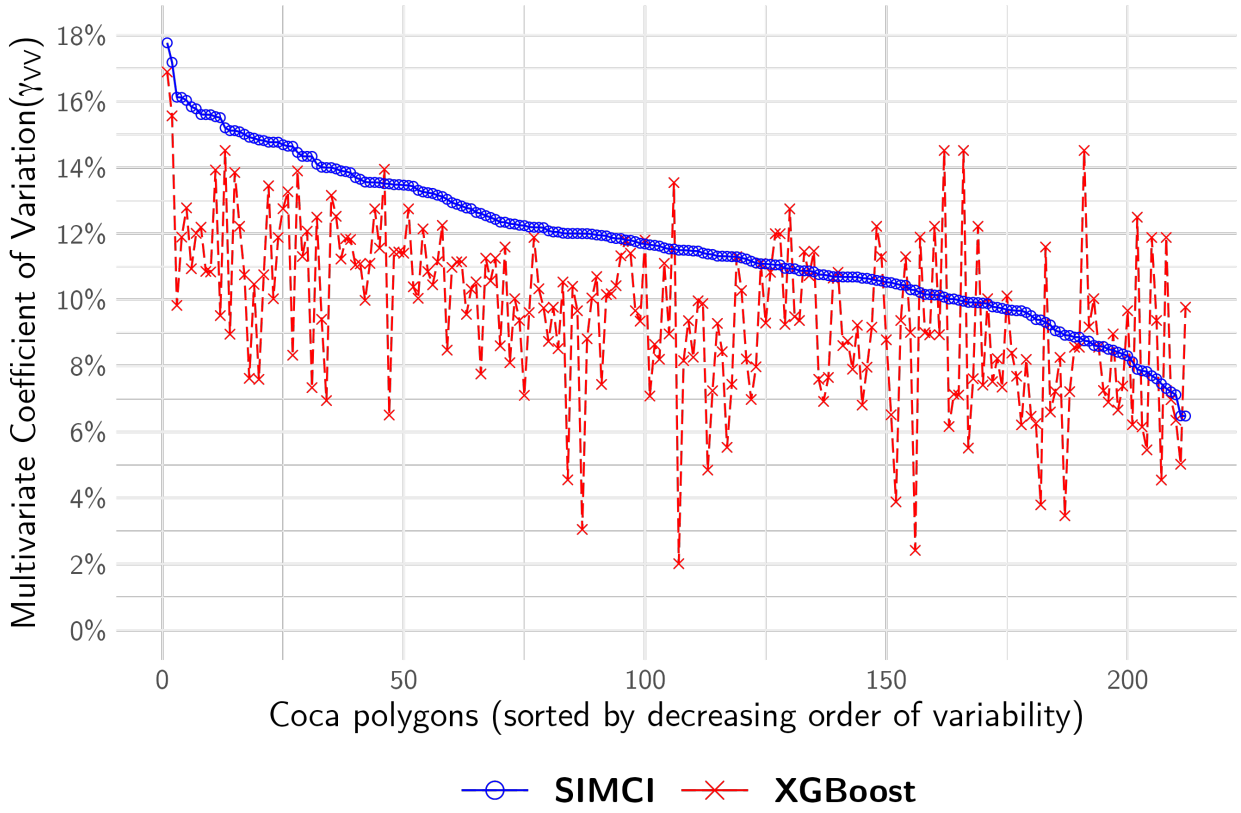

Figure 4-9: MCV values for SIMCI polygons and corresponding model polygons. Points below the blue line indicate instances where the model's polygon has a lower MCV than the corresponding SIMCI polygon
Source: Author's own

and delineate coca crops with a high degree of geometric precision. However, the model's performance does not exist in isolation—it's inherently affected by the characteristics of the reference data, such as the distribution and configuration of the polygons.

For instance, the model demonstrates substantial precision in cases where the boundaries of coca crops are distinct and well-defined. However, in complex scenarios where multiple polygons are clustered together, the model's delineation may fall short of the reference data. This suggests that the model may struggle with accurately capturing the geometry of coca crops in more challenging and complex scenarios.

4.2.5. Variable importance

The analysis of variable importance provides valuable insights into the contributions of each feature to the predictive power of the model. For the second dataset, we considered

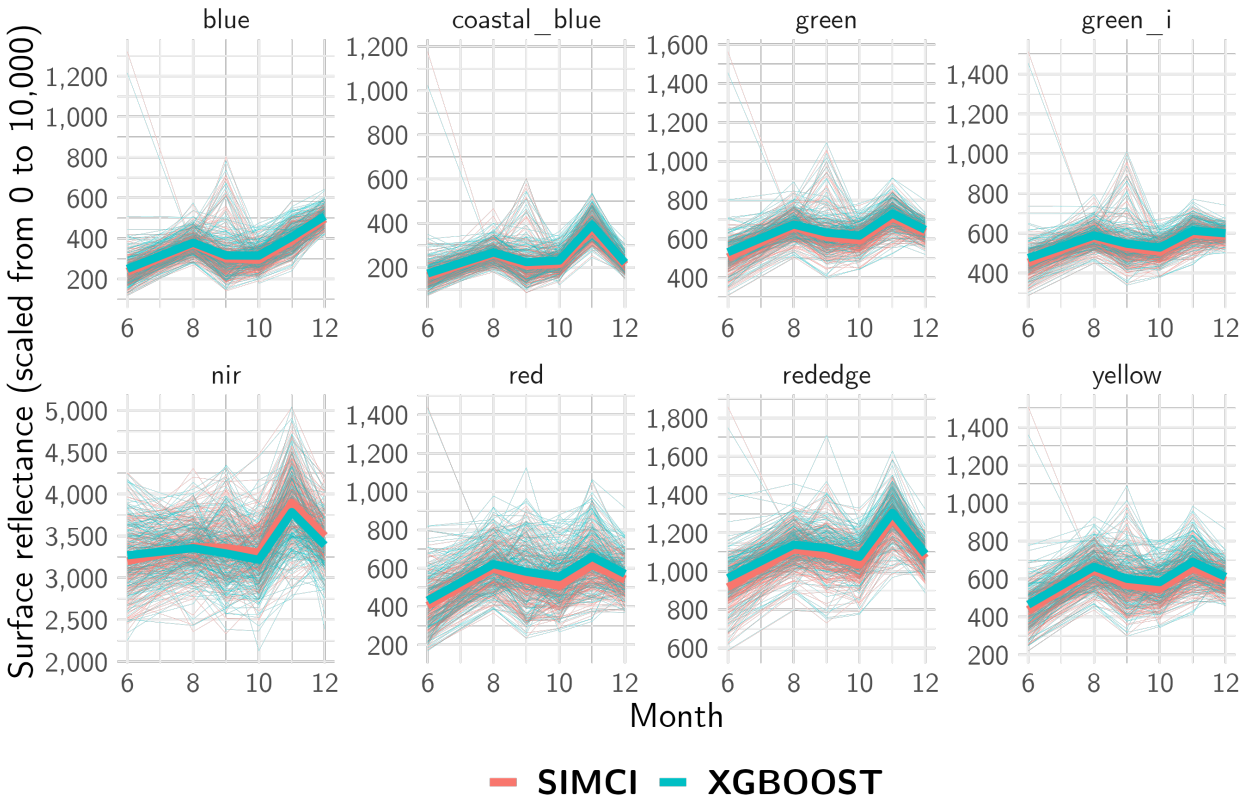

Figure 4-10: Temporal profiles for the eight bands for both SIMCI and XGBoost coca polygons

Source: Author's own

nearly 200 features, encompassing basic bands and their focal forms. The 20 most influential features, as determined by their contribution to the model's gain, were primarily in their focal forms.

Particularly, two features, 'diciembre_rededge_mean' and 'diciembre_rededge_max', stood out due to their significantly higher contribution to the model's performance. These two features were predominantly associated with the red-edge band of the December month's imagery. The prominence of these variables highlights the importance of the red-edge band, renowned for its ability to capture vegetation characteristics, and the significance of the December imagery, which may correspond to a specific stage in the coca growth cycle. This observation is in line with prior research in coca classification [Ángel, 2012], which examined the role of the red edge parameter in distinguishing different stages of coca growth.

This finding suggests that the model was heavily influenced by the variations in the red-edge spectral characteristics of the coca crops in December. It not only provides valuable insights into the spectral-temporal features most critical to the model's performance but also indicates

the potential to improve model efficiency by focusing on key variables. The list of these 20 variables, ranked by their importance, can be seen in Figure **4-11**.

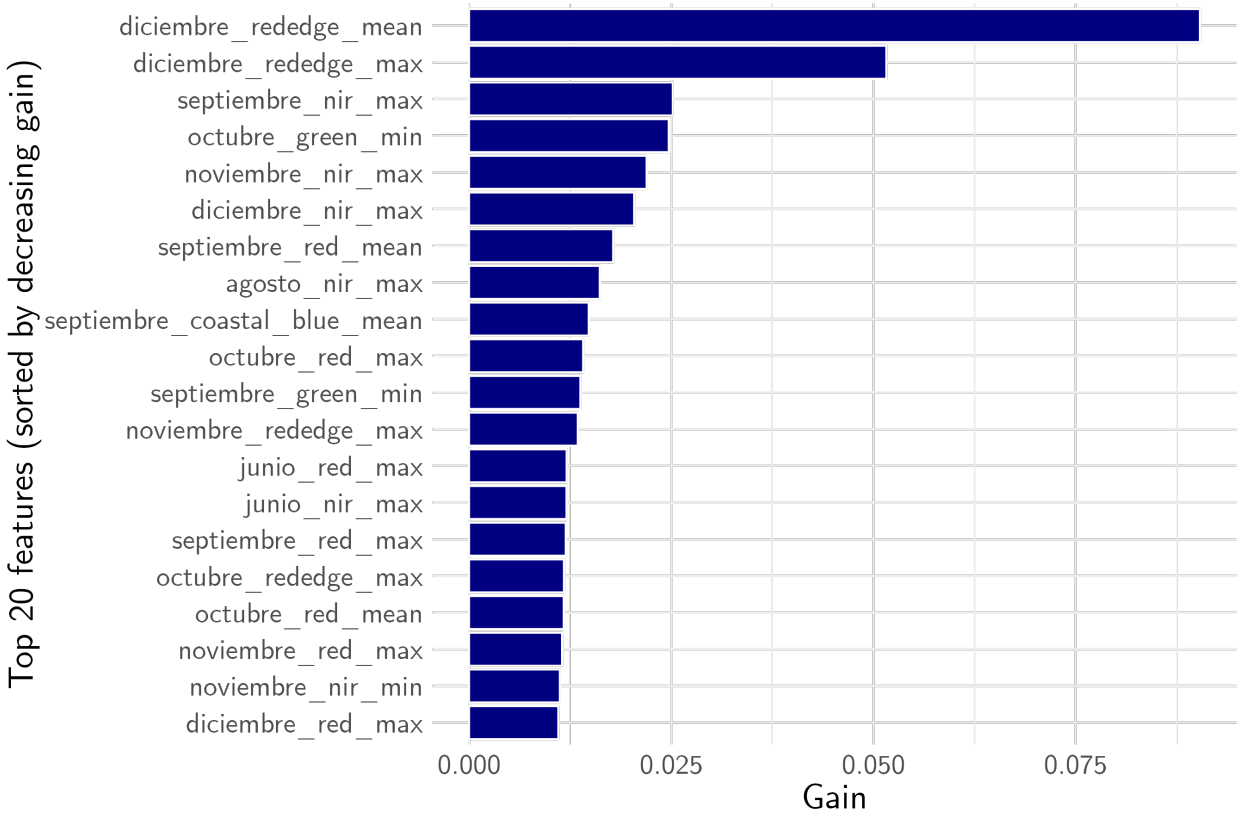

Figure 4-11: Variable importance of top 20 features for the Dataset 2 best model, evaluated by the Gain

Source: Author's own

5 Conclusions and future work

5.1. Conclusions

This research explored how machine learning can contribute monitor coca crops in Colombia using satellite imagery. While coca cultivation in the country presents numerous challenges due to its large geographic extent, intricate nature, and associated sociaeconomic complexities, this study provides initial insights into potential solutions to facilitate the monitoring of the crop. With the utilization of Sentinel-2 and Planet imagery, and the application of the Extreme Gradient Boosting (XGBoost) algorithm, the research aimed to present a robust, replicable, and scalable method for coca crop monitoring.

The study's outcomes shed light on the potential of machine learning in contributing significantly to understanding and managing the coca cultivation landscape. The following sections delve into the detailed conclusions derived from this research, discussing the effectiveness of machine learning, the significance of specific spectral-temporal features, and the importance of spatial resolution and cloud cover in the context of the study. The conclusions also underscore the areas of future research to further refine and expand upon the current methodology.

5.1.1. Effectiveness of Machine Learning

The application of machine learning methods, specifically the Extreme Gradient Boosting (XGBoost) algorithm, for the task of coca crop monitoring in Colombia has proven to be highly effective within the study's geographical scope. The research focused on the Putumayo region, covering two zones over a span of three years, providing a solid foundation for a machine learning approach in this specific context.

Our methodology provides a robust and repeatable quantitative approach to satellite image analysis, improving upon traditional, subjective interpretations.

One of the significant advantages of the machine learning approach is its scalability. While the current study was limited to the Putumayo region, the developed model can potentially be extended to other geographical areas, given appropriate adjustments and validations.

Furthermore, the model's speed and efficiency make it a powerful tool for dealing with the vast geographical scope of coca cultivation.

Moreover, the proposed machine learning approach enables automation of the process, significantly reducing the time and effort required for large-scale analysis. Given the vast volume and high temporal resolution of modern satellite imagery, manual analysis becomes increasingly challenging, if not infeasible. Implementing an automated, machine learning-based method can manage these large datasets effectively and provide frequent, consistent updates, aiding in timely decision-making for coca crop management.

5.1.2. Superior Performance

Our study has demonstrated the superior performance of an XGBoost-based model in coca crop classification. Notably, the Kappa score achieved in this study, 0.7512, represents a significant improvement over the score of 0.709 reported by [Ángel, 2012]. This increased precision speaks volumes about the potential of machine learning in improving the monitoring of coca crops.

This enhancement in performance can be primarily attributed to the use of novel, comprehensive feature sets, advanced machine learning methods, and the exploitation of spatial and temporal information. However, it's important to remember that this performance was evaluated only in specific geographic locations (Putumayo region, two zones) and over a certain time period (three years). Therefore, while the results are undoubtedly encouraging, further research is needed to validate these findings across different geographical locations and over extended periods of time.

Our approach not only performed exceptionally well but also set a new standard (state-of-the-art) in coca crop classification, with an Overall Accuracy (OA) of 94.3 % and 93.5 % in Dataset 1 and Dataset 2, respectively. As per the available public studies and data, this work represents a breakthrough, underscoring the model's potential for accurate and reliable monitoring of coca cultivation. This substantial progress can greatly bolster the creation of more efficient counter-narcotics strategies. This accomplishment represents a significant stride in the field, setting a new benchmark, with the challenge now being to continue advancing and elevating this standard.

5.1.3. Important Features

In the course of the study, it was discovered that the SWIR 1, SWIR 2 bands in the Sentinel-2 imagery and the Red Edge band in the Planet imagery, along with their spatial and temporal versions, were consistently significant for discerning between coca and not coca classes. These findings underline the importance of these specific spectral bands and their spatial features in the context of coca cultivation identification.

The temporal aspect also played a crucial role in the model's performance. The features corresponding to the last months of the year, particularly December, demonstrated remarkable importance. This underscores the ability of the model to adapt to the spectral-temporal dynamics of the coca growth cycle, indicating that the later quarter of the year may be particularly relevant for detecting coca cultivation.

Interestingly, spectral indices, commonly used in remote sensing, didn't improve the model's predictive power. This suggests that decision tree-based models, like XGBoost, are capable of capturing complex relationships between spectral bands, rendering spectral indices unnecessary and even redundant.

On a more practical note, the findings imply that the model can potentially be implemented with different sensors, as long as they provide the necessary bands or equivalents. In other words, the model's efficacy is not bound to specific sensor platforms, which greatly enhances its flexibility and applicability.

These conclusions serve as a valuable guide for future studies, indicating that focus should be placed on harnessing specific bands and their spatial and temporal features, and less on spectral indices. This can help in designing more efficient models for coca crop detection, opening avenues for scalable, accurate, and robust monitoring systems.

5.1.4. Improved Geometric Precision with High-Resolution Imagery

The use of high-resolution imagery from the Planet sensor enabled a more precise geometric representation of the coca crops. Despite the lower spectral resolution of Planet images, they were able to contribute to an improved model performance in Dataset 2. The improved geometric precision in identifying and delineating the coca crops was particularly advantageous in detecting smaller plots and understanding the complex, often fragmented, nature of coca cultivation landscapes.

This led to a more accurate and realistic comprehension of the spatial distribution of coca crops, which can be crucial for both strategic decision-making and operational actions on

the ground. For instance, the ability to identify smaller coca plots can aid in prioritizing areas for eradication efforts, and understanding the spatial patterns of cultivation can contribute to the development of more effective anti-narcotics strategies.

The results highlight the potential advantages of integrating high-resolution imagery into the model, even at the cost of reduced spectral information. It suggests that future studies can benefit from further exploring and optimizing the use of high-resolution sensors in the machine learning-based coca classification tasks.

5.1.5. Need for Further Research

While this study provides a promising foundation for the use of machine learning methods in coca crop monitoring in Colombia, it remains a preliminary investigation within a limited scope - two zones in the Putumayo department over three years, and a binary classification model.

Further research should aim to increase the model's complexity and sophistication to capture a wider variety of land cover and land use types. This could include incorporating more classes and subclasses, testing the model's adaptability across different regions, and integrating other remote sensing data types such as synthetic aperture radar (SAR) imagery. Exploring alternative machine learning methods, including deep learning techniques, and different validation sources, such as eradication lots, would also offer valuable insights.

Ultimately, these prospective research directions underscore the complexity of the coca cultivation issue in Colombia. The battle against coca cultivation extends beyond technical and scientific realms, requiring a holistic approach that encompasses social, economic, political, and environmental dimensions. This study contributes one piece to this multifaceted puzzle, pointing towards innovative, efficient, and effective strategies to address this persistent problem.

5.2. Future work

This research represents a pioneering effort to develop a data-driven, statistically robust method for monitoring coca crops in Colombia. The approach leverages advanced machine learning algorithms, multi-spectral satellite imagery, and geospatial data to deliver promising results in terms of scalability, efficiency, and accuracy. However, as with any novel undertaking, there is ample room for refinement, expansion, and exploration of additional aspects. The future work directions identified here aim to build upon the foundation laid

by this study, address potential shortcomings, and further enhance the model's performance and applicability.

5.2.1. Inclusion of more classes and subclasses

The current model primarily focuses on distinguishing between coca crops and not coca areas. However, the complexity and diversity of land use and crop types extend beyond this binary classification. Therefore, introducing more nuanced classes and subclasses can provide a more comprehensive understanding of the landscape and improve the model's precision. Potential expansions could include:

- Within the 'coca' class: Differentiating between varying densities of coca plantations, identifying associations with other crops, or even attempting to categorize different varieties of the coca plant, if feasible.
- Within the 'not coca' class: Distinguishing among different land uses such as forest, other crop types, water bodies, built-up areas, etc.

By capturing a more granular representation of the land use and crop types, the model could provide more actionable insights for crop monitoring and management.

5.2.2. Testing other sensors

As this study relied mainly on Sentinel-2 and Planet imagery, future work could involve testing the model with data from other remote sensing technologies, to compare and evaluate their efficacy. Incorporating data from varied sensor types could potentially enhance the model's robustness and precision. Here are a few possibilities:

- *Other Optical Sensors:* Though Sentinel-2 and Planet provided a balance between spatial resolution and spectral bands, there is still room to explore imagery from other optical sensors. The high-resolution sensors, such as those onboard WorldView or GeoEye satellites, might offer improved geometric precision. However, these sensors typically come with fewer spectral bands compared to Sentinel-2, which might influence the model's ability to distinguish coca crops based on their spectral characteristics. Nonetheless, the potential impact of these sensors in enhancing the model's performance warrants further investigation.
- *Radar Technology:* Synthetic Aperture Radar (SAR) sensors, such as those onboard Sentinel-1 or RADARSAT Constellation, could be another potential data source. Unlike optical sensors, radar can penetrate cloud cover, making it a reliable data source in regions with persistent cloudiness. Additionally, radar backscatter can provide information about surface roughness and structure, which may provide additional distinguishing features for coca crops.

It is also important to consider the trade-offs associated with these different sensor types, such as the balance between spatial and spectral resolution, cost of data, and susceptibility to weather conditions.

5.2.3. Application to other regions

While the model has demonstrated effectiveness in the Putumayo context, testing its applicability to other regions would provide insights into its scalability and transferability. This would help validate the model's robustness in diverse geographical and environmental contexts.

5.2.4. Exploring other methods

While XGBoost has proven effective for this task, there's a wide spectrum of machine learning algorithms to explore. Implementing and comparing results from methods like neural networks could provide interesting insights and potentially enhance the model's performance.

5.2.5. Incorporating other ground truths

The use of additional ground truth datasets, such as eradication plots, could supplement the validation process and potentially enhance the model's predictive accuracy. A more diverse ground truth can help mitigate bias and improve the generalizability of the model.

References

[Abdikan et al., 2023] Abdikan, S., Sekertekin, A., Narin, O. G., Delen, A., and Sanli, F. B. (2023). A comparative analysis of SLR, MLR, ANN, XGBoost and CNN for crop height estimation of sunflower using sentinel-1 and sentinel-2. *Advances in Space Research*, 71(7):3045–3059.

[Aerts et al., 2015] Aerts, S., Haesbroeck, G., and Ruwet, C. (2015). Multivariate coefficients of variation: Comparison and influence functions. *Journal of Multivariate Analysis*, 142:183–198.

[Arbia, 2014] Arbia, G. (2014). *A Primer for Spatial Econometrics*. Palgrave Macmillan UK.

[Aybar et al., 2020] Aybar, C., Wu, Q., Bautista, L., Yali, R., and Barja, A. (2020). rgee: An r package for interacting with google earth engine. *Journal of Open Source Software*, 5(51):2272.

[Bohorquez et al., 2017] Bohorquez, M., Giraldo, R., and Mateu, J. (2017). Multivariate functional random fields: prediction and optimal sampling. *Stochastic Environmental Research and Risk Assessment*, 31(1):53–70.

[Bouhennache et al., 2018] Bouhennache, R., Bouden, T., Taleb-Ahmed, A., and Cheddad, A. (2018). A new spectral index for the extraction of built-up land features from landsat 8 satellite imagery. *Geocarto International*, 34(14):1531–1551.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

[Breiman et al., 1984] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification And Regression Trees*. Routledge.

[Ceballos and Lopera, 2009] Ceballos, N. and Lopera, G. (2009). El caso coca nasa. *Cuadernos de Investigación*.

[Ceccato et al., 2002] Ceccato, P., Gobron, N., Flasse, S., Pinty, B., and Tarantola, S. (2002). Designing a spectral index to estimate vegetation water content from remote sensing data: Part 1. *Remote Sensing of Environment*, 82(2-3):188–197.

[Chen et al., 2004] Chen, D., Stow, D. A., and Gong, P. (2004). Examining the effect of spatial resolution and texture window size on classification accuracy: an urban environment case. *International Journal of Remote Sensing*, 25(11):2177–2192.

[Chen and Guestrin, 2016] Chen, T. and Guestrin, C. (2016). Xgboost: Scalable tree boosting system. *CoRR*, abs/1603.02754.

[El Financiero, 2023] El Financiero (2023). Destruyen casi 18,000 plantas de coca en guatemala.

[El Tiempo, 2023] El Tiempo (2023). La coca florece en méxico a la sombra de drogas sintéticas.

[ESA, nda] ESA, E. S. A. (n.d.a). User guides - sentinel-2 msi - overview - sentinel online. https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-2-msi/overview. Accessed: 2023-07-29.

[ESA, ndb] ESA, E. S. A. (n.d.b). User guides - sentinel-2 msi - resolutions - sentinel online. https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-2-msi/resolutions. Accessed: 2023-07-29.

[ESA, ndc] ESA, E. S. A. (n.d.c). User guides - sentinel-2 msi - revisit and coverage - sentinel online. https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-2-msi/revisit-coverage. Accessed: 2023-07-29.

[Farfán and Jaramillo, 2009] Farfán, F. and Jaramillo, A. (2009). Sombrío para el cultivo del café según la nubosidad de la región. *Avances Técnicos*.

[Frampton et al., 2013] Frampton, W. J., Dash, J., Watmough, G., and Milton, E. J. (2013). Evaluating the capabilities of sentinel-2 for quantitative estimation of biophysical variables in vegetation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 82:83–92.

[Friedman et al., 2000] Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, 28(2).

[Friedman, 2001] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5).

[Galindo and Fernández, 2010] Galindo, A. and Fernández, J. (2010). Plantas de coca en colombia. discusión crítica sobre la taxonomía de las especies cultivadas del género erythroxylum p. browne (erythroxylaceae).

[Gerber et al., 2018] Gerber, F., de Jong, R., Schaepman, M. E., Schaepman-Strub, G., and Furrer, R. (2018). Predicting missing values in spatio-temporal remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 56(5):2841–2853.

[Haboudane, 2004] Haboudane, D. (2004). Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: Modeling and validation in the context of precision agriculture. *Remote Sensing of Environment*, 90(3):337–352.

[Huang et al., 2022] Huang, L., Liu, Y., Huang, W., Dong, Y., Ma, H., Wu, K., and Guo, A. (2022). Combining random forest and XGBoost methods in detecting early and mid-term winter wheat stripe rust using canopy level hyperspectral measurements. *Agriculture*, 12(1):74.

[Huber et al., 2022] Huber, F., Yushchenko, A., Stratmann, B., and Steinhage, V. (2022). Extreme gradient boosting for yield estimation compared with deep learning approaches. *Computers and Electronics in Agriculture*, 202:107346.

[Huete, 1988] Huete, A. (1988). A soil-adjusted vegetation index (SAVI). *Remote Sensing of Environment*, 25(3):295–309.

[InfoBAE, 2022] InfoBAE (2022). Autoridades localizan 25.000 arbustos de hoja de coca en el caribe hondureño.

[Jensen, 2006] Jensen, J. R. (2006). *Remote Sensing of the Environment*. Prentice Hall.

[Jiang and Shekhar, 2017] Jiang, Z. and Shekhar, S. (2017). *Spatial Big Data Science*. Springer International Publishing.

[Karpatne et al., 2016] Karpatne, A., Jiang, Z., Vatsavai, R. R., Shekhar, S., and Kumar, V. (2016). Monitoring land-cover changes: A machine-learning perspective. *IEEE Geoscience and Remote Sensing Magazine*, 4(2):8–21.

[Louhaichi et al., 2001] Louhaichi, M., Borman, M. M., and Johnson, D. E. (2001). Spatially located platform and aerial photography for documentation of grazing impacts on wheat. *Geocarto International*, 16(1):65–70.

[Machado et al., 2019] Machado, M. R., Karray, S., and de Sousa, I. T. (2019). LightGBM: an effective decision tree gradient boosting method to predict customer loyalty in the finance industry. In *2019 14th International Conference on Computer Science &amp Education (ICCSE)*. IEEE.

[Matteucci and Morello, 2001] Matteucci, S. and Morello, J. (2001). Aspectos ecológicos del cultivo de coca. 1(8).

[McFEETERS, 1996] McFEETERS, S. K. (1996). The use of the normalized difference water index (NDWI) in the delineation of open water features. *International Journal of Remote Sensing*, 17(7):1425–1432.

[Mishra and Mishra, 2012] Mishra, S. and Mishra, D. R. (2012). Normalized difference chlorophyll index: A novel model for remote estimation of chlorophyll-a concentration in turbid productive waters. *Remote Sensing of Environment*, 117:394–406.

[Mohanaiah et al., 2013] Mohanaiah, P., P.Sathyanarayana, and GuruKumar, L. (2013). Image texture feature extraction using glcm approach. *International Journal of Scientific and Research Publications*, 3(5).

[Mutanga and Kumar, 2019] Mutanga, O. and Kumar, L. (2019). Google earth engine applications. *Remote Sensing*, 11(5):591.

[Nazir et al., 2021] Nazir, A., Ullah, S., Saqib, Z. A., Abbas, A., Ali, A., Iqbal, M. S., Hussain, K., Shakir, M., Shah, M., and Butt, M. U. (2021). Estimation and forecasting of rice yield using phenology-based algorithm and linear regression model on sentinel-II satellite data. *Agriculture*, 11(10):1026.

[Oliver and Webster, 1990] Oliver, M. A. and Webster, R. (1990). Kriging: a method of interpolation for geographical information systems. *International journal of geographical information systems*, 4(3):313–332.

[Park et al., 2021] Park, J., Lee, Y., and Lee, J. (2021). Assessment of machine learning algorithms for land cover classification using remotely sensed data. *Sensors and Materials*, 33(11):3885.

[Park et al., 2018] Park, S., Im, J., Park, S., Yoo, C., Han, H., and Rhee, J. (2018). Classification and mapping of paddy rice by combining landsat and SAR time series data. *Remote Sensing*, 10(3):447.

[PBC, 2022] PBC, P. L. (2022). Combined imagery product specifications.

[Piedelobo et al., 2019] Piedelobo, L., Hernández-López, D., Ballesteros, R., Chakhar, A., Pozo, S. D., González-Aguilera, D., and Moreno, M. A. (2019). Scalable pixel-based crop classification combining sentinel-2 and landsat-8 data time series: Case study of the duero river basin. *Agricultural Systems*, 171:36–50.

[Rouse et al., 1974] Rouse, J. W., Haas, R. H., Schell, J. A., Deering, D. W., et al. (1974). Monitoring vegetation systems in the great plains with erts. *NASA Spec. Publ*, 351(1):309.

[Rustowicz, 2017] Rustowicz, R. M. (2017). Crop classification with multi-temporal satellite imagery.

[Saini and Ghosh, 2021] Saini, R. and Ghosh, S. K. (2021). Crop classification in a heterogeneous agricultural environment using ensemble classifiers and single-date sentinel-2a imagery. *Geocarto International*, 36(19):2141–2159.

[Serrano, 2014] Serrano, M. (2014). Cultivos ilícitos de coca y bienestar en las regiones productoras: Un análisis desde el enfoque de capacidades.

[Tatsumi et al., 2015] Tatsumi, K., Yamashiki, Y., Torres, M. A. C., and Taipe, C. L. R. (2015). Crop classification of upland fields using random forest of time-series landsat 7 ETM data. *Computers and Electronics in Agriculture*, 115:171–179.

[Tobler, 1970] Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46:234.

[UNODC, 2023] UNODC (2023). World drug report 2023.

[UNODC and MINJUSTICIA, 2010] UNODC and MINJUSTICIA (2010). Características agroculturales de los cultivos de coca 2005-2010.

[UNODC and MINJUSTICIA, 2012] UNODC and MINJUSTICIA (2012). Estructura económica de la unidades productores agropecuarias en zonas de influencia de cultivos de coca.

[UNODC and MINJUSTICIA, 2019] UNODC and MINJUSTICIA (2019). Monitoreo de cultivos ilícitos 2018.

[UNODC and MINJUSTICIA, 2022] UNODC and MINJUSTICIA (2022). Monitoreo de cultivos ilícitos 2021.

[Valavi et al., 2018] Valavi, R., Elith, J., Lahoz-Monfort, J. J., and Guillera-Arroita, G. (2018). blockcv: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution*, 10(2):225–232.

[Valen, 1974] Valen, L. V. (1974). Multivariate structural statistics in natural history. *Journal of Theoretical Biology*, 45(1):235–247.

[Yu and Zhu, 2020] Yu, T. and Zhu, H. (2020). Hyper-parameter optimization: A review of algorithms and applications.

[Ángel, 2012] Ángel, Y. (2012). Metodología para identificar cultivos de coca mediante análisis de parámetros red edge y espectroscopia de imágenes.