Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis Double Degree in Econometrics/Economics

Satellite-Based Location of Cocoa Farms in Ghana and Côte d'Ivoire: Trends, Implications, and EU Deforestation Law

Dmitriy Knyazhitskiy (523327)

Supervisor:	Andrea Naghi
Second assessor:	Stan Koobs
Date final version:	01 July 2023

Abstract

Locating cocoa farms in Western Africa is a challenging task. Cocoa farms are of small size, no public location data is available and it is difficult to visually distinguish cocoa agroforestry from tropical forest. The satellite vegetation imagery of Ghana and the Côte d'Ivoire are used to classify cocoa farms on a pixel level. The cocoa probability map created by Kalischek et al. (2022) is used for creating labels and the XGBoost classifier is trained on the multidimensional satellite images achieving a classification accuracy of 94.5%. Our results suggest that the months of December and January are most suitable for separating cocoa land cover. The annual cocoa production is estimated across different years and we observe a significant correlation with available production data. The obtained annual cocoa farms location maps seem to be aligned with the existing research. The cocoa farms change map highlighting the changes in the cocoa farms' locations in the last 10 years is presented. New farms appearing around the border between Ghana and Côte d'Ivoire are detected. The implications of the EU deforestation law are discussed and we advocate for using satellite imagery for locating (illegal) farms in protected areas requiring additional support due to the EU deforestation law.

Contents

Replication part

This research was inspired by Khachiyan et al. (2021). The authors show that using high-resolution satellite imagery in the context of deep learning can create accurate forecasts of GDP. In this paper, we develop a different application of satellite imagery and apply it to different parts of the world. Yet, the objective stays the same - create insights that are relevant to the well-being of certain groups of people. As discussed with my supervisor, the replication part may be skipped if the extension is elaborate enough and no code/insight is used from the original paper. Hence, replication is not included in this thesis.

1 Introduction

1.1 Cocoa production overview

Cocoa beans are one of the most traded agricultural commodities in the world. The primary use of cocoa beans is the production of chocolate. Other uses include producing cocoa powder or beverages. According to the Centre for the Promotion of Imports from Developing Countries (2022), the world's average chocolate consumption is 0.9kg/year. The country with the highest chocolate consumption per capita in 2022 was Germany with 11kg/year followed by Switzerland with 9.7kg/year. With the EU chocolate market valued at €42 Billion in 2022, Europe is the largest importer of cocoa beans worldwide. In the 2021/2022 season Africa was responsible for 81% of global cocoa beans production with 43% of production in Côte d'Ivoire and 20% in Ghana, together these two regions account for more than 60% of global cocoa production (Kakaoplattform, 2022). The total EU market value of Ghana, Ivory Coast and Cameroon was €4.6 billion in 2021 (European Commission, 2022). The area potentially suitable for cocoa production in these regions is around 300 000 km^2 which is approximately equal to the area of Italy.

Most of the chocolate farms are of a size of one to four hectares (equivalent to one to six football fields) and are operated by a single family that lives nearby. On average, the chocolate output is about 400kg/ha (Wessel & Quist-Wessel, 2015). This organisation is quite different from other farmed commodities that can be planted on large and easily observable fields. Small cocoa beans farms spread all over Ghana and Ivory Coast are often undocumented and are challenging to locate.

There is no accurate data about cocoa production available which complicates the analysis of the cocoa industry and poses challenges for understanding its impact. The cocoa industry appears to be adverse both internally (for local producers and inhabitants) and externally (for the rest of the world). A lot of cocoa farmers live below the poverty line and many cases of slavery and child labour are documented (Food Empowerment Project, 2022). Cocoa production contributes to deforestation as the tropical forest has to be cut in order to build new farms. This also leads to substantial CO_2 emissions. Multiple protected areas and national parks exist to preserve nature and wildlife, but illegal farms are located in these regions (Higonnet, Bellantonio & Hurowitz, 2017). With the EU deforestation law, the cocoa landscape in Western Africa might change resulting in job loss for some farmers. Therefore, this research aims to utilize satellite

imagery to locate cocoa farms, estimate the annual crop area and production, understand where the new farms are appearing and discuss the impact of the EU deforestation law on the local population. The presented analysis can be used as a basis for policies directed at helping the vulnerable groups of people in Ghana and the Ivory Coast.

To demonstrate the lack of consensus about cocoa production numbers, we present two resources of production data. One is provided by the Food and Agriculture Organization of the United Nations (2023) and the other one is provided by Statista (2023). Figure 1 shows that the data from these sources are different. Although the scale of data varies a lot, these two series exhibit a significant correlation of 0.77. In order to evaluate our production estimates against available data, a measure that captures the relationship between variables while allowing for the variation of scale is required. A correlation coefficient is picked as such a measure to evaluate our predictions.

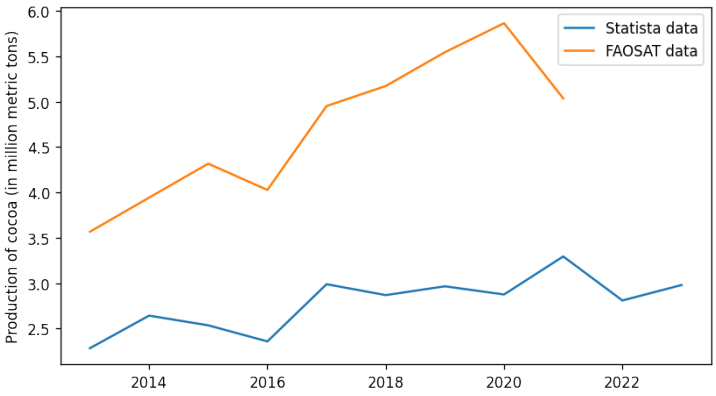

Figure 1: Production data from different sources.

Cocoa production maps are required for training cocoa farms' classification models. We found two cocoa production maps of Ghana and the Ivory Coast where the cocoa farms are highlighted. Kalischek et al. (2022) provides one of these maps and another one is provided by the European Comission (2023). Although both use satellite imagery, they arrive at quite different results. Figure 2 demonstrates the cocoa maps around Kumasi, a large city in Ghana. We see the substantial difference between the two images. The map created by the European Comission (2023) appears to be less reliable, as it seems to create homogeneous predictions (so the located farms' spatial density does not vary a lot) which is questionable given the number of different land covers and the presence of a variety of protected areas and national parks (Abu, Szantoi, Brink, Robuchon & Thiel, 2021).

1.2 Deforestation overview

Deforestation is one of the largest global environmental concerns. It not only distorts wildlife, but it is also responsible for CO_2 emissions and hence global warming via the greenhouse effect[1].

It is estimated that 10 million hectares of forest were cut down each year (Food and Agriculture Organization of the United Nations, 2020). The world lost one-third of its total forest in the last 10,000 years. One-half of total forest loss happened in the 20th and 21st centuries

[1]Gases like carbon dioxide trap the heat inside an atmosphere preventing it from escaping to space.

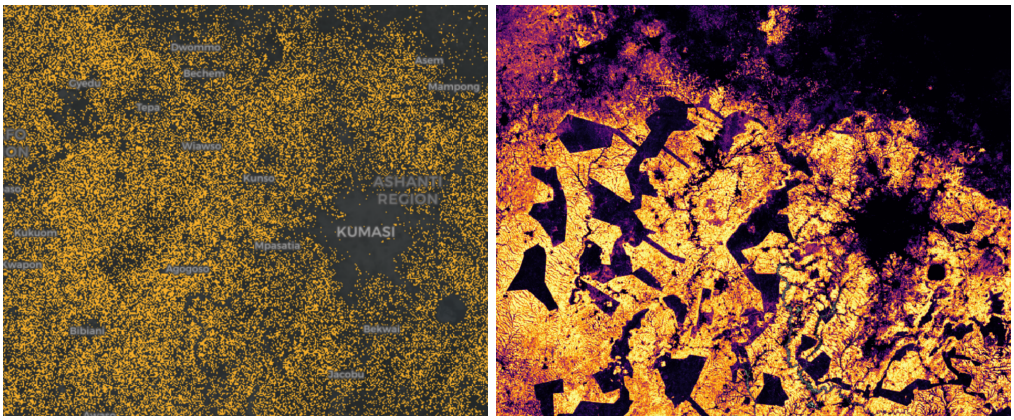

Figure 2: Map comparison, European Comission (2023) (left) and Kalischek et al. (2022) (right). Areas marked yellow show cocoa farms, while dark-coloured areas mean no cocoa farms are detected.

with the area lost equal to the area of the United States (Ritchie & Roser, 2021). Pendrill et al. (2019) estimate that tropical deforestation is responsible for 6.5% of global CO_2 emissions. For comparison, the CO_2 emissions from the civil aviation industry are responsible for 'only' 2.5% of global emissions (Quadros, Snellen, Sun & Dedoussi, 2022).

This forest loss is mostly driven by agricultural demands for new crop areas and one-quarter of global forest loss comes from tropical forest loss. Ritchie and Roser (2021) advocate that tropical forest deforestation is the most harmful way of deforestation and stopping it should be prioritised. Not only due to geographical considerations but also because it harms biodiversity a lot; more than half of global species reside in tropical forests and when the forest is cut down these species might lose their habitats permanently (Scheffers, Joppa, Pimm & Laurance, 2012). In the case of cocoa beans, animals that suffer the most are elephants, chimpanzees, crocodiles, pygmy hippos and leopards (Higonnet et al., 2017).

With the growth of the human population, an increase in agricultural production is expected. Yet, this does not have to lead to increased deforestation because for most commodities the crop yields were consistently growing (see Figure 6 for example with cocoa). The crop yield is defined as a ratio of production to total crop area. For the last sixty years, the total amount of land for agriculture increased only by 7% while the global population more than doubled (Higonnet et al., 2017). Currently, most deforestation happens in South America, Central Africa, Southeast Asia and Western Africa, however, deforestation affects the entire planet.

By purchasing products produced in these deforested regions, developed countries including Europe and US contribute to the deforestation of tropical forests. Moreover, by creating long-term demand for such products, these countries give an additional incentive for the farmers to destroy more tropical forests. The EU deforestation law adopted on the 16th of May of 2023 will oblige EU companies to make their supply chains deforestation-free. This law will have a substantial impact on the cocoa industry and will be discussed more extensively later in this paper.

1.3 Existing research

Studying cocoa farm locations using remote sensing became popular in the last five years. It is possible to consider raw imagery as input, but a more advanced method is to transform the RGB[2] data into spectral signatures. Spectral signature refers to the intensity of surface reflection for various bands. A band is a range of wavelengths corresponding to a certain spectrum, it is a relevant concept in our study. There is also qualitative research analysing the general trends.

Kalischek et al. (2022) perform extensive research on static (non-time series) cocoa farms' location in Ghana and the Ivory Coast. The authors collected more than 10,000 cocoa farm locations from various sources and they train a deep neural network to make the cocoa probability map. They used nine bands from Sentinel-2 imagery with a spatial resolution of ten meters and they achieve 85% classification accuracy on the part of the same imagery not used for training. Figure 3 demonstrates the probability map of cocoa plantations the authors create. It will be used as a 'ground truth' dataset in our research.

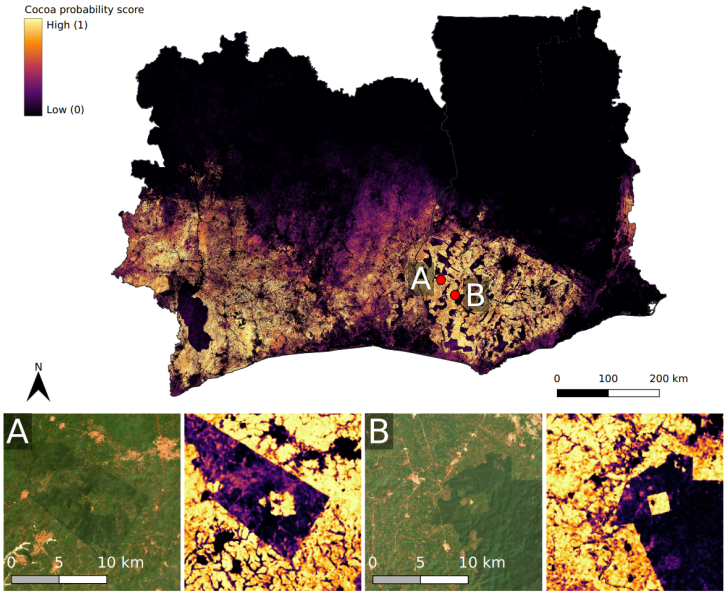

Figure 3: Cocoa farm probability map, taken from Kalischek et al. (2022).

Batista et al. (2022) perform more theoretical research studying spectral signatures of cocoa agroforest. Compared to the Kalischek et al. (2022), they provide more insight into their methodology and they first analyse what bands can be potentially useful for locating cocoa farms by investigating their spectral signatures. They show that cocoa has a spectral signature highly similar to tropical forest which makes cocoa forest separation a challenging task. They also note that some bands, such as 8a (Red 864.8nm) in some months can be used to separate cocoa from forest. The authors use multiple approaches such as Ridge, Random Forest and XGBoost classifiers to make binary predictions. Again, the data is trained and evaluated on different parts of the same multidimensional pictures. In this study, the XGBoost classifier will be used and the details can be found in the Methodology Section.

[2]Red-Green-Blue (RGB) is a way to encode a colour as a 3-dimensional array representing the weights of corresponding colours.

Although the mentioned research provides tools for a deep understanding of cocoa farms' locations, it does not consider the farm location dynamics with time. Moreover, there is no evidence such research will be robust when presented models are applied to another image. Satellite imagery is 'cleaned' and standardised, but there are still some variations in the spectral range quality of images and pixel locations across different time periods. On the other hand, survey analysis performed by Yao Sadaiou Sabas, Gislain Danmo, Akoua Tamia Madeleine and Bogaert (2020) does consider the time series dynamics of cocoa production. The authors analyse the cocoa production in Ivory Coast between 1985 and 2015. Instead of satellite imagery, their data is based on surveys and it is helpful to understand the trends of the age population and cocoa farm sizes. However, no predictions or estimations about the cropland size of the entire country can be made from such analysis due to the imprecision of the collected data and delays (as these surveys spread for a significant amount of time). Hence, our research will also contribute to the existing literature by using remote sensing to analyse the changes in cocoa farms' locations over time.

1.4 Potential challenges

There are four main contributors to the complexity of locating cocoa farms.

First, the available satellite imagery resolution is not high. The Vegetation satellite data is available in the minimal spatial resolution[3] of 250 meters. This already exceeds the size of many cocoa farms meaning that the typical farm will consist of less than one pixel on such an image.

The photographic satellite imagery data is available in a higher resolution (up to 10-meter spatial resolution for Sentinel-2 imagery). But this brings the second complication - memory and computational complexity required for building a successful classification model. Batista et al. (2022) used spectral signatures of cocoa beans to predict the location of cocoa farms. They required almost one terabyte of RAM to perform their analysis. Although it is possible to perform similar research on this scale using cloud services, we prefer to use more computationally feasible solutions.

In case the Machine Learning classification model is implemented, a very limited amount of labelled data (farm or not farm) is available. Often, the used datasets are not supplied by researchers as cocoa farm data is manually collected and labelled requiring a considerable amount of effort, hence this data is kept as proprietary. Filella (2018) used only fifteen farms in their analysis for cocoa farms location. Although this analysis demonstrated good accuracy for detecting large cocoa farms, it misses out on the detection of smaller farms and their model would require new training and labelling for every new satellite image. Batista et al. (2022) applied the model trained in Brazil to the analysis of Ghana and the Ivory Coast. We will instead use the cocoa map created by Kalischek et al. (2022) - it appears as the most reliable publicly available dataset of cocoa farms (predicted) locations. This poses a limitation - we use labels created by authors as 'ground truth' data. If the author's labels are imprecise, it will be a challenge to our research.

Finally, the cocoa farms' spectral signature is very similar to the other surfaces, especially to the forest. This similarity is demonstrated in Batista et al. (2022) and is also confirmed by our

[3]Spatial resolution corresponds to a real-world size of 1 pixel of an image.

findings. Figure 4 shows the spectral signature of the cocoa farms and everything else, where 'everything else' refers to any area in Ghana and the Ivory Coast that was not classified as cocoa (forest, city, lakes, desert, etc). Separating the two might indeed be a challenging task meaning that the forest cover satellite imagery such as presented in Higonnet et al. (2017) might actually understate the deforestation effect by misclassifying the cocoa farms as forest.

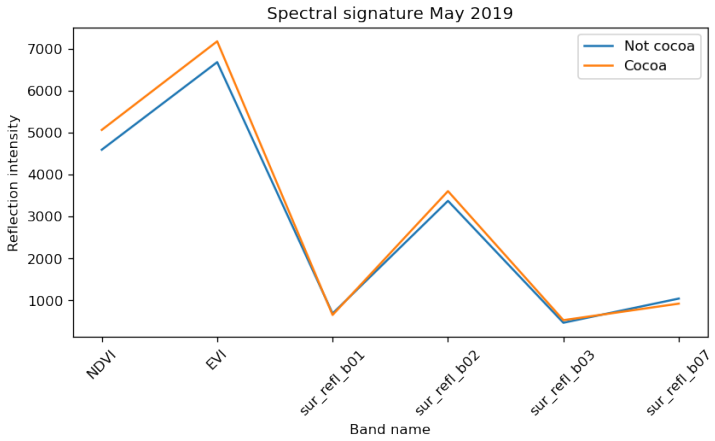

Figure 4: Spectral signature of cocoa farms in May 2019, average for pixels with labels 'cocoa' and 'not cocoa'. The x-axis represents various bands - which will be clarified in the Data Section below, and the y-axis represents reflection intensity.

2 Data

We use satellite imagery provided by Google Earth Engine and two sources of annual production data. The most challenging part of this research was not building models, rather it was collecting and manipulating data, finding ways to effectively store and compress it as well as selecting appropriate cloud servers for processing.

2.1 Is crop area related to production?

The annual data for cocoa beans production and crop area for both Ghana and Ivory Coast can be accessed by the Food and Agriculture Organization of the United Nations (2023). It is not easy to find data at a higher frequency than annually. Moreover, the data in a higher frequency might be imprecise as there are only two main crops of cocoa beans per year and the exact collection time might differ among farms. The data is available for 1994-2021, but some values are missing and they were imputed by the data providers. As it can be seen in Figure 5, cocoa production was increasing consistently since 1994.

There is a strong relationship between crop area and production and it is evident from crop yield analysis. Figure 6 shows that after the year 2000, there is a clear growth in crop yield, caused by better agriculture practices and an increase in fertiliser usage (Ruf, Schroth & Doffangui, 2015).

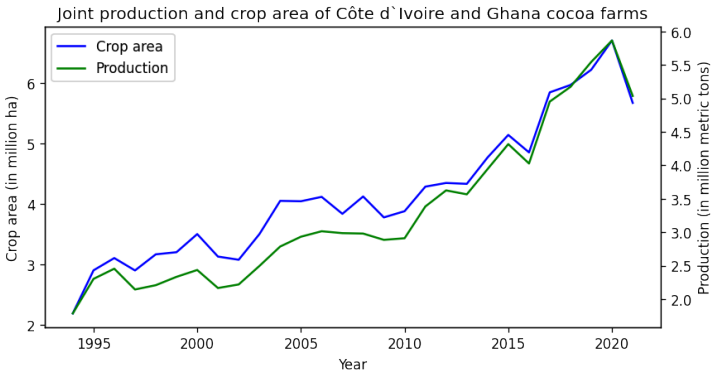

Figure 5: Production and crop area, 1994-2021.

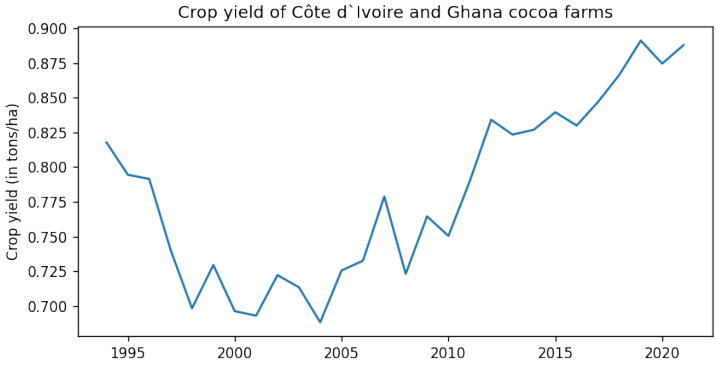

Figure 6: Crop yield, 1994-2021.

2.2 Satellite imagery

Satellite vegetation imagery of a region of about 500,000 km^2 (12 times the size of the Netherlands) is used. Usually, similar analyses are performed using Sentinel-2 imagery (Abu et al., 2021; Batista et al., 2022). Sentinel's mission is to detect changes in land cover and to monitor the world's forests. It was launched in 2015 and offers a spatial resolution of up to ten meters.

Yet, we decided to use the MODIS (Moderate Resolution Imaging Spectroradiometer) satellite. This dataset, accessed through the Google Earth Engine consists of global vegetation images created by Google Earth Engine Team (2023). There were two main reasons we chose this satellite. First, MODIS imagery is calibrated over time which makes a time series imagery comparison much easier. Second, the spatial resolution of this imagery is 250 meters, which is on the one hand not optimal, but on the other hand, it allows to feasibly process a much larger amount of data (it now takes $(250/10)^2 = 625$ times less space compared to Sentinel).

MODIS data is available on a wide variety of bands, the two main bands are NDVI (Normalized Difference Vegetation Index) and EVI (Enhanced Vegetation Index). Both bands are created from other raw satellite bands to enhance the differences between land covers. The EVI band performs better in high biomass conditions. Other available bands are sur_refl_b01 (Red surface reflectance), sur_refl_b02 (Near-infrared[4] surface reflectance), sur_refl_b03 (Blue surface

[4]Near infrared is light of invisible to human eye wavelength, but still close to visible light, the wavelength is 858nm. See Appendix B for illustration.

reflectance), sur_refl_b07 (Mid-Infrared[5] surface reflectance) and some other bands that are not used in this research. Images are already cleaned from an atmosphere colour, clouds and cloud shadows.

2.3 Dataset for model training

The cocoa probability map is reprojected onto the monthly satellite imagery of 2019 to obtain labels for every pixel. A classifier is trained on a small subset of the available data.

Kalischek et al. (2022) created the aggregate cocoa map for the years 2017 to 2020, but because we are interested in one-year forecasts, we will assume the map is correct for 2019. We will train the classifier with their labels for the year 2019 and then apply this classifier to different years. It is also possible to train four classifiers for each year separately and then use an average prediction.

We use 12 monthly images throughout 2019 as a training set. For every image we use the cocoa probability map from Kalischek et al. (2022) and project[6] it on the MODIS vegetation imagery. The authors suggest that the optimal classification threshold is 0.65, meaning that if the predicted probability is below 0.65, a pixel is classified as 'not cocoa'. We use this map and the suggested threshold to create binary labels of 'cocoa' and 'not cocoa'. This way two images are obtained, one with cocoa farms only and another with no farms, see Figure 7. This image is created for every band and for every month, so in total, the 72 dimensions for each label are analysed.

Not surprisingly, there are more labels 0 (not cocoa) than 1 (cocoa). Imbalanced classifiers are not favourable as our focus is primarily on the minority class (cocoa) and imbalanced classifiers might cause the predictions to be biased in favour of the majority class (not cocoa). Luckily, we have a very large dataset and we will rely on the undersampling technique. Undersampling is a technique to balance skewed datasets by using only a part of the majority class data (Hastie, Tibshirani & Friedman, 2001).

We split our area of interest in grids of 0.15^o x 0.15^o = 16.7 km x 16.7 km squares (with about 4,400 pixels per square). Each square might have a mix of labels 0 and 1 or even no labels at all if this square falls into the Atlantic Ocean. We shuffle the grid and keep the order of permutation the same for all analyses. Then for squares in a grid, we iterate through each pixel and record pixel coordinates, pixel spectral signature and label. We stop after the count of both label 0 and label 1 exceeds 1,000,000.

It is important to ensure that across different months one row of the final dataset (one pixel) has the same location. Otherwise, it is possible that values for different locations are recorded together. This is avoided by matching the coordinates of the pixels in 12 spectral signatures datasets. Luckily for us, in the MODIS dataset, pixels generally are at the same location over time. However, the images are taken from a physical satellite and some imprecisions exist. Hence, we will identify pixels 'the same' if the deviation across both coordinates is less than $0.0001^o = 11$ meters. In other words, we round the pixel coordinates up to the fourth digit and use them as a key for joining the tables in different months.

[5]Reflectance with a wavelength of 2130nm, not visible to a human eye. See Appendix B for illustration.

[6]Here projecting refers to rescaling the pixels of both images to the same scale while preserving the coordinate system.

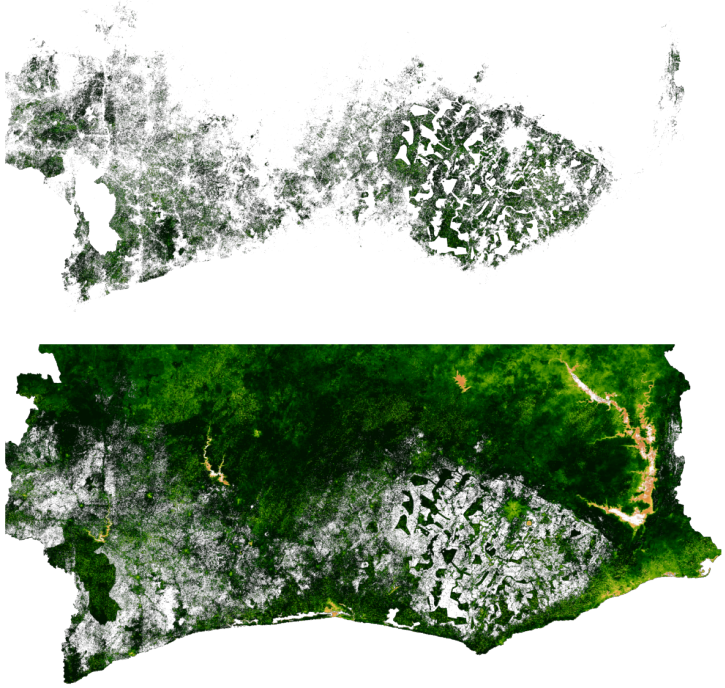

Figure 7: NDVI band imagery of Ghana and Ivory Coast coastline where cocoa farms are located (top) and not located (bottom).

The final dataset consists of about 1,950,000 rows and is about 1 GB in size. We randomly sample 80% of the data for training and retain the remaining 20% for testing our model.

2.4 Dataset for time series predictions

We collect the data for different years in order to make predictions. The procedure is similar to the one described above, we create a grid of points, iterate through it saving data and use a model to predict labels. In this case, we still don't use the entire dataset, instead, about 4,400,000 pixels are sampled as approximately 1000 of 0.15^o x 0.15^o squares. Figure 8 shows the region that will be analysed in this study.

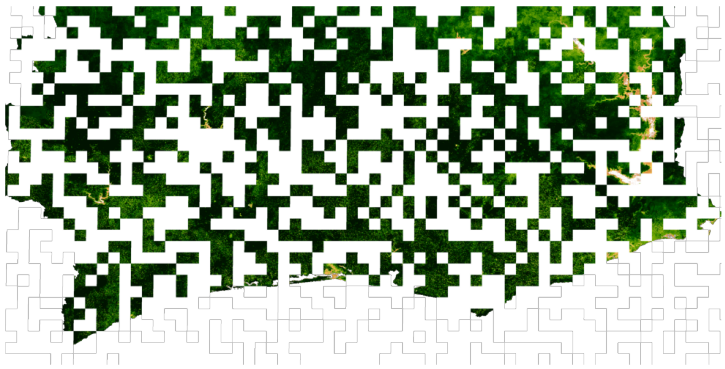

Figure 8: Area used for prediction of cocoa production across years, shown in NDVI band.

This dataset takes approximately 3 GB per year and it is of course possible to use the entire sample to obtain a more precise forecast. Yet, we doubt that the increase in accuracy will be

significant, while the time needed to collect the data might increase substantially. Time plays an important role because the response speed of the Earth Engine API is limited. It takes about six hours to collect the selected data subsample for one year, although this process can be multithreaded by creating several API instances.

3 Methodology

To analyse trends and estimate cocoa crop area, an accurate pixel-level classifier is needed. Section 3.1 justifies the use XGBoost classifier, Section 3.2 further describes the relevant concepts from XGBoost and Section 3.3 explains performed hyperparameter optimisation. In Section 3.4 the selected classifier is discussed, the confusion matrix is provided and results are compared to classifiers from other research. Section 3.5 describes how to convert predicted probability arrays into crop area estimates and calibration of the predicted probability curve. In Section 3.6 we take a deeper dive into the classifier and the optimal classification threshold is selected. Section 3.7 demonstrates selected cocoa density maps and the methodology for obtaining cocoa change maps is described.

3.1 Method Selection

We evaluate different machine learning methods and justify the use of the XGBoost classifier for our problem.

Usually, deep learning and in particular the Convolutional Neural Networks (CNN) are used for extracting insights from satellite images. Filella (2018) and Kalischek et al. (2022) are using deep learning for cocoa classification. Abu et al. (2021) are using Random Forests and Batista et al. (2022) are using a variety of methods including Random Forest, Ridge Regression and XGBoost and they conclude XGBoost seems to be the best classifier.

The main advantage of CNN is their convolutional layer which allows it to learn complex spatial patterns of the objects (Wu, 2017). In other words, the network is 'aware' of the object's shape, which is advantageous for high-resolution satellite imagery as it will be able to recognise cocoa fields rather than classify each pixel inside a field separately. Yet, in our case, the farms are already of a size of a single pixel and hence there is no strong spatial dependency in the data (of course, some regions will have a higher density of farms, but these are, most likely, different farms rather than one big farm so it makes no sense to aggregate these points). This also means that our data can be represented as an unordered tabular dataset (a collection of rows and columns where the order of rows can be arbitrary) without a substantial loss of information.

There is a long ongoing debate about whether the Boosted Trees or Neural Networks perform better for tabular datasets. This can be seen from the outcomes of Kaggle competitions [7] (Goldbloom, 2016). There is also academic research arguing that both methods perform similarly and hyperparameters optimisation is more important than the model architecture selection (McElfresh et al., 2023). The authors also show that to achieve similar performance, Neural Networks took 1-2 orders of magnitude longer training time. Therefore, we would also prefer to use XGBoost as it will allow us to explore more parameters.

[7] Kaggle is a large platform where various data science competitions are held.

3.2 XGBoost

We describe relevant concepts (with examples) from classification trees, bagging, random forest and boosting and link them to XGBoost parameters.

XGBoost, which stands for eXtreme Gradient Boosting, was introduced by Microsoft and it quickly became one of the most popular boosted trees methods (Chen & Guestrin, 2016). As of June 2023, the paper which introduced this method is cited more than 27,000 times. XGBoost is an ensemble method, meaning that it combines multiple weak predictors into one strong predictor. Boosted trees are similar to random forests, yet while random forests are a simple average of the decision trees, boosted trees are built on top of each other. Among other things, boosted trees are able to put more emphasis on misclassified instances. On the other hand, boosted trees require more parameter tuning.

In order to understand XGBoost, it helps to understand the basics of tree-based methods. The simplest method is a decision tree. Decision trees — in particular, classification trees — are non-parametric methods used for supervised learning. The core idea is to split a sample region into many rectangular subsamples and fit a simple model (as simple as a constant) for each rectangle (Hastie et al., 2001). In order to build a decision tree, the algorithm uses a top-down approach, recursively selecting a way to split data based on certain criteria. During the tree construction, the algorithm iterates through features and possible splits for each node searching for the best split. The 'best' split is a split that minimises a certain loss function (impurity measure), such as Gini impurity or cross-entropy. The goal of a loss function is to calculate the dissimilarity measure between the predicted and true probabilities in each class. One of the most popular loss functions is cross entropy. It calculates the dissimilarity as an average number of bits needed to represent true labels given the predicted probabilities. For a node m, a number of observations in that node N_m and a selected region R_m in that node, the cross-entropy of a binary classifier is defined as

$$-\hat{p}_m \log(\hat{p}_m) - (1 - \hat{p}_m) \log(1 - \hat{p}_m), \tag{1}$$

where

$$\hat{p}_m = \frac{1}{N_m} \sum_{x_i \in R_m} I\left(y_i = 1\right)$$

$\hat{p}_m$ is a fraction of observation labelled as 1. Figure 9 shows an example decision tree constructed for a small subsample of our dataset where the number of terminal nodes (final nodes) is restricted by 10. Inside each box there is a final split, Gini impurity measure, amount of points in each subsample, number of points for each class and current class label, these values provide an intuition behind the classification tree mechanism.

Although the decision trees are fast and allow for model visualisation, they suffer from many disadvantages including overfitting, sensitivity to the input data (as the tree is built top to bottom, a small change in a dataset might result in a substantially different tree) and high variance. Bootstrap aggregating (bagging) introduced by Breiman (1996) helps to tackle this issue by creating multiple decision trees via bootstrapping the training dataset. Bootstrapping is a range of techniques focused on generating artificial data by sampling with replacement from the original dataset. The predictions are done using a 'majority voting', so if more trees predict a

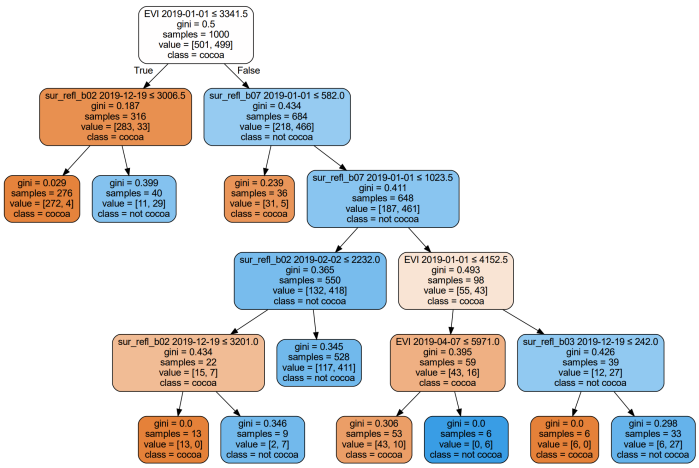

Figure 9: Example of decision tree for a subset of cocoa dataset.

label 1, this label is assigned to a datapoint as a prediction. Bagging allows to decrease variance and improves the robustness of the results at the cost of interoperability as now it's not possible to represent the model as a single tree (Hastie et al., 2001). It will help to think of bagging as introducing diversity between training data.

The next significant improvement in ensemble methods was the introduction of random forests (Breiman, 2001). The bagged trees are highly correlated, as they are trained on similar datasets, hence the variance of final predictions does not decrease as much as one might expect. Random forests aim to decrease the correlation between trees and hence achieve an even higher reduction in variance. Imagine, for example, in our dataset, the EVI band in May appears to be by far the most significant feature. In that case, each tree in the bagging algorithm is likely to only select the EVI May feature in every split, meaning that other less significant features are neglected and some information is lost. While this is acceptable for regression methods (selecting only a subset of features with the highest impact) there is no need to do the same in the ensemble trees framework as 'weak' models can be effectively combined with 'stronger' models. Random forests achieve this by only considering a fraction of features at every split. This way random forest help to provide a decrease in variance and generally it achieves a higher performance and is easier to tune as the increasing number of trees does not overfit the model. It is possible to calculate the feature importance in the random forest by looking at the total information gain change for a given feature at each split. Figure 10 illustrates the scaled feature importance graph obtained.

In boosting, unlike in random forests where all trees are separate, trees are built sequentially, which allows adjusting the weight of each tree in an ensemble. Thus, more difficult cases can receive a higher weight. Gradient boosting is a special kind of boosting where boosting errors are minimised using the gradient descent (see Appendix C) algorithm. Note that contrary to random forests, increasing the number of trees might yield overfitting.

XGBoost is a state-of-the-art gradient boosting algorithm. Among other things, XGBoost is optimised for parallel processing and compatible with GPUs, it also allows for the regularisation of loss functions and is able to handle missing values. Equation 2 shows the XGBoost loss

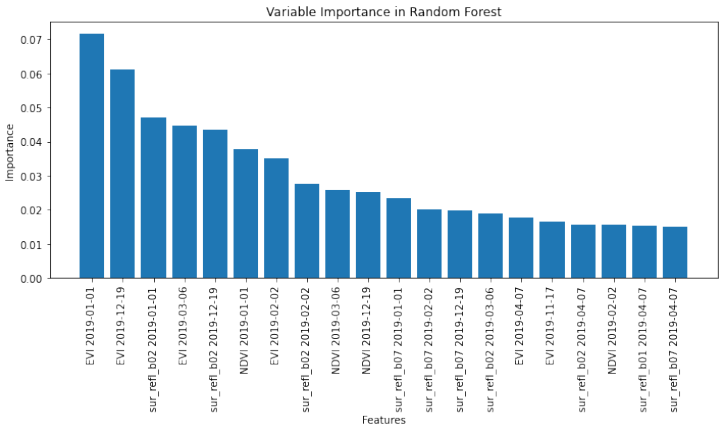

Figure 10: Variable importance for random forest.

function, where l is a differentiable convex loss function (binary cross entropy in our case) and Ω is a convex regularisation consisting of penalty for a number of leaves in a tree T and L_1 and L_2 penalties for leaf weights (Chen & Guestrin, 2016).

$$\mathcal{L}(\phi) = \sum_i l\left(\hat{y}_i, y_i\right) + \sum_k \Omega\left(f_k\right), \tag{2}$$

where

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|_2^2 + \alpha \|w\|_1$$

Equation 3 shows the nature of sequential training. If $\hat{y}_i^{(t)}$ is the prediction of the i-th instance at the t-th iteration, the idea is to add f_t to minimise the remaining unexplained part.

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t\left(\mathbf{x}_i\right)\right) + \Omega\left(f_t\right) \tag{3}$$

Afterwards, the second-order expansion similar to the one explained in Appendix C is used to find optimal split points.

There are multiple parameters that can be adjusted in XGBoost models. The first group of parameters are general parameters such as 'learning rate', 'number of estimators', 'subsample' and 'colsample bytree'. The learning rate is a rate of the gradient descent algorithm explained in Appendix C. Higher learning rate is generally associated with a lower number of estimators. When the learning rate is high each tree captures more information, but this also means that if the number of estimators is high, overfitting is more likely as the model memorises more special cases. 'Subsample' is a fraction of the dataset used to train each tree, this is similar to bagging. 'Colsample bytree' is a fraction of features used for training each tree, it is a similar concept to random forest column sampling. However, while random forests perform a different sampling at each leaf, the XGboost does it on a tree level.

The second group are Tree-specific parameters and they include 'max depth', 'reg alpha' and 'reg lambda'. The 'max depth' parameter controls the maximum allowed depth of a tree. For example, the tree in Figure 9 has a depth of 5. Note that the number of leaves grows exponentially with the depth, a tree of depth n can have 2^{n+1} leaves. 'Reg alpha' and 'reg

lambda' are regularisation terms applied to weights of features as shown in equation 2. These coefficients have a standard interpretation of L_1 and L_2 penalty terms.

3.3 Hyperparameter optimisation

We use a cloud server with NVIDIA RTX4000 GPU (graphics processing unit) for training our models. By utilising GPUs, XGBoost is able to process thousands of computations in parallel. We see approximately a ten times increase in processing speed compared to training on twelve Intel Xeon E5 core CPUs (Central Processing Units). The total training time took about 24 hours, given that most modern laptops have four to six cores, the training on such laptops will take twenty to thirty days.

In order to select the best model, we initialise a grid of classifier parameters and then perform the random search parameter optimisation with three-fold cross-validation. Random search parameter optimisation involves sampling random parameters from a parameter grid. It is considered to be better than the grid search parameter optimisation (where all points on the grid are used) as it allows to cover of wider ranges of points in a more sparse way (Bergstra & Bengio, 2012). In total, we sample 272 points from the parameter grid.

The cross-validation is performed by separating the data into three folds (groups) of equal size, training the model on two folds and evaluating the performance on the remaining fold. This process is then repeated two more times using a different validation fold each time. After that, the average score is recorded as the final score for this set of parameters. This process helps to increase the robustness of the results and to decrease the variance of estimates at the price of the computational time.

We use a binary log-loss as a loss function (see equation 1) and we use classification accuracy as an evaluation metric because our dataset is balanced. Table 1 shows a grid of parameters we used.

Table 1: Hyperparameter Values.

Hyperparameter	**Value**
max_depth	(4, 6, 8, 12, 20, 25)
learning_rate	(0.01, 0.05, 0.1, 0.2, 0.3)
subsample	(0.6, 0.8, 1)
colsample_bytree	(0.6, 0.8, 1.0)
reg_lambda	exp(-12,-11,..5)
reg_alpha	exp(-12,-11,..5)
random_state	0
n_estimators	100

We can also see how the training accuracy evolved across different parameter values. Figure 11 shows the test score distribution for a given parameter value. For instance, the left top panel shows how scores were distributed when the max_depth was 5, 10, 15, etc. We see that the most improvement happens from increasing the max_depth parameter. However, this parameter influences the computational time substantially, when we set the parameter max_depth to 25, the resulting tree has $2^{26} - 1 = 67108863$ nodes and it does not seem feasible to increase it even further.

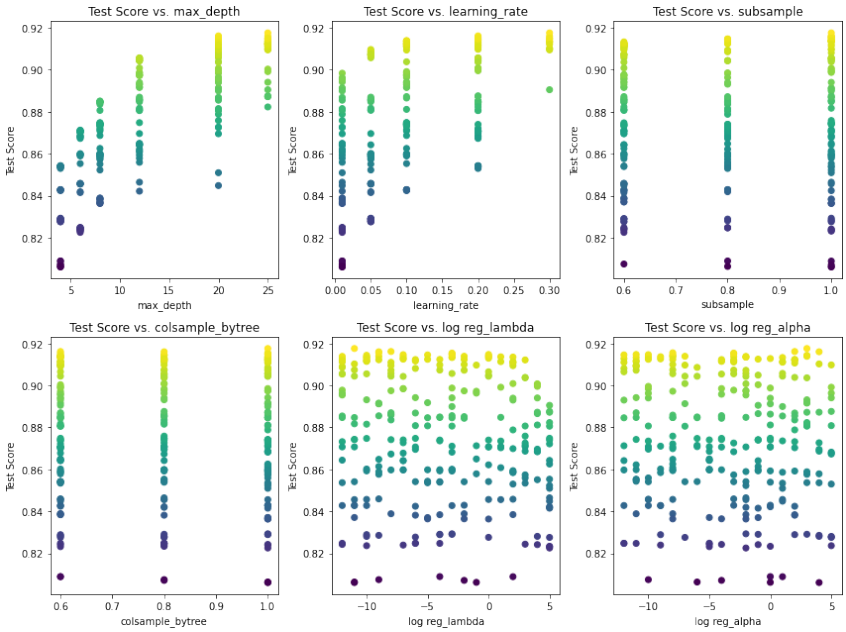

Figure 11: Score vs parameter values. Every graph demonstrates cross-validation scores when one parameter value is fixed.

3.4 Classifier

We selected our best classifier via cross-validation. Table 2 shows the optimal parameters selected. With this classifier and a default classification threshold of 50%, we have achieved an out-of-sample accuracy of 92.5%. The confusion matrix shown at Table 3 shows that the number of false negatives exceeds the number of false positives, so some bias in predictions exists even for a balanced dataset.

Table 2: Best hyperparameters.

Hyperparameter	**Value**
max_depth	25
learning_rate	0.3
subsample	1
colsample_bytree	1
reg_lambda	20.1
reg_alpha	1.67
random_state	0
n_estimators	100

Table 3: Confusion Matrix for classification threshold of 0.5.

	Actual 0	**Actual 1**
Predicted 0	177,248	19,048
Predicted 1	10,677	189,064

Accuracy is not a sufficient metric used for the evaluation of the classifier, it is also helpful to understand whether the error rate is driven by false positives. Typically, the two additional

measures used are precision and recall. In remote sensing problems, the term User Accuracy (UA) is used for precision and the term Producer Accuracy (PA) is used for recall. Thus, User Accuracy shows how many of the images predicted as cocoa are actually cocoa, while Producer Accuracy shows how many images with the true label cocoa are correctly predicted.

Abu et al. (2021) achieved UA of 62.2% and PA of 82.9%. However, we should keep in mind that these are predictions for the entire countries of Ghana and Ivory Coast and the fraction of cocoa farms is about 8%. Batista et al. (2022) reports the highest accuracy out of 7 models of 95.2% with UA and PA of 94.3% and 90.8% respectively. Their training set is much smaller and consists of approximately 25% labels. Our baseline paper (Kalischek et al., 2022) achieved an accuracy of 85.9% with UA and PA 88.5% and 87.2% accuracy. Finally, our classifier achieved an accuracy of 92.5%, UA of 94.7% and PA of 90.8% on a balanced dataset. Our results appear to be quite good compared to the existing research.

3.5 Classification threshold

The classifier was trained on the balanced dataset, which is not representative of the fraction of cocoa agroforest. In order to adjust our classifier we will resample a representative dataset (so a dataset where the fraction of cocoa farms reflects the fraction for the entire dataset) and vary a classification threshold. By threshold, we mean such value p that given a continuous prediction $\hat{p}$ we label a datapoint as 1 if and only if $\hat{p} > p$. We do this instead of directly training a model on a representative dataset because we are interested in the cocoa, which is a minority class and we want to ensure there is no bias towards the majority class. The standard accuracy metrics might be misleading for imbalanced datasets. Usually in remote sensing classifications oversampling (increasing the size of the minority class by generating artificial data) is used (Douzas, Bacao, Fonseca & Khudinyan, 2019). Oversampling is driven by the same objective to remove class imbalance, but it is not feasible computationally for our research.

The data from the Food and Agriculture Organization of the United Nations (2023) suggest that about 8% of the total area of Ghana and Ivory Coast is labelled as cocoa, while for Statista (2023) it's closer to 4%. In our labelled dataset, this fraction is 17%. Given that our area of interest covers about half of the total area of Ghana and Ivory Coast and we anticipate no cocoa farms outside of our area of interest we get a similar fraction to the Food and Agriculture Organization of the United Nations (2023).

Using that fraction of cocoa farms is about 17%, we sample a representative dataset from the test set and arrive at the dataset consisting of 237240 points with a fraction of 1 labels approximately equal to 0.17. Because we will vary the classification threshold and it can potentially yield overfitting, we also split the resulting test set into 2 parts, one with 189792 points for classification analysis and another with 47448 points to access and validate the final results.

Left panel 12 shows the false positive and false negative rate as a function of threshold. The false positive rate is defined as $FPR = \frac{FP}{FP+TN}$ and the false negative rate is defined as $FNR = \frac{FN}{FN+TP}$. This graph can provide us with an intuition of where an optimal threshold can be. The middle panel of Figure 12 demonstrates the receiver operating characteristic (ROC) curve. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) for various threshold values and is a general representation of a binary classifier. For a random

prediction, we would expect a straight line and the further the ROC curve is from the line, the better the classifier. In our case, the classifier is distant from the 45-degree line, which is not unexpected given the achieved prediction accuracy, meaning that a wide range of thresholds can be considered. Finally, the right panel shows prediction accuracy as a function of the threshold. The highest accuracy is achieved for the threshold of 0.90. It is not surprising this number exceeds 0.5, we only label cocoa as 'cocoa' if our confidence level is high, so cases when we are uncertain fall under more likely 'not cocoa' case.

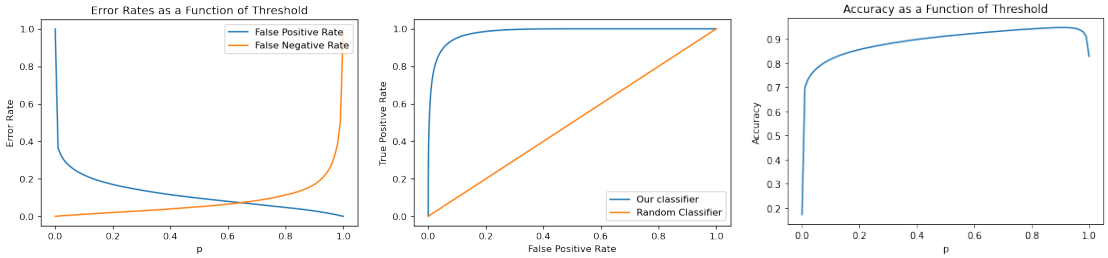

Figure 12: ROC curve (left), FP and FN rate (middle) and accuracy as function of threshold (right).

3.6 Prediction setup

In order to find an estimate of the crop area, we need to find a way to aggregate the predicted probabilities of each pixel into a single number representing the fraction of the total area. The problem of finding this fraction can be considered as follows: There exists an unknown 1 x n array X with binary random variables and a probability estimate array $\hat{X}$ consisting of $\hat{p_i}$ for each element. We want to know the expected fraction of variables labelled 1. As long as we assume the $\hat{p}$ is an unbiased estimator of p - true probability of being 1, the answer is straightforward: $E(\sum_{i=1}^{n} x_i) = \sum_{i=1}^{n} E(x_i) = \sum_{i=1}^{n} E(p_i) = \sum_{i=1}^{n} \hat{p_i}$. Hence, we can take an average probability across the input array as a scale-invariant proxy for the crop area.

Is $\hat{p}$ an unbiased estimator of p? Generally, it is not guaranteed. A way to access the quality of predictions is to create a calibration plot (reliability curve). These curves allow to access the quality of classifiers by plotting the true frequency of predicted labels for intervals of predicted probability. For example, when the predicted probability is between 60% and 70% we would expect the fraction of positive labels around 65% for an unbiased classifier. As it can be seen in Figure 13, although some deviations exist, they are relatively not substantial. Calibrating the predicted probability did not yield better results either.

We also need a simple estimate for a yield curve and for this, we fit a linear trend model $y_t = c + a * (t - 2000)$ for t = 2000...2021 where y_t is crop yield demonstrated in Figure 6. This should allow us to get a more robust estimate of the yield curve and cocoa production.

3.7 Change in cocoa farms' location

By representing the density of farms on the map and then running a regression for each sub-region we can gain insight into the dynamics of farm location changes.

Predictions of cocoa probability for every pixel can be obtained. However, this will be difficult

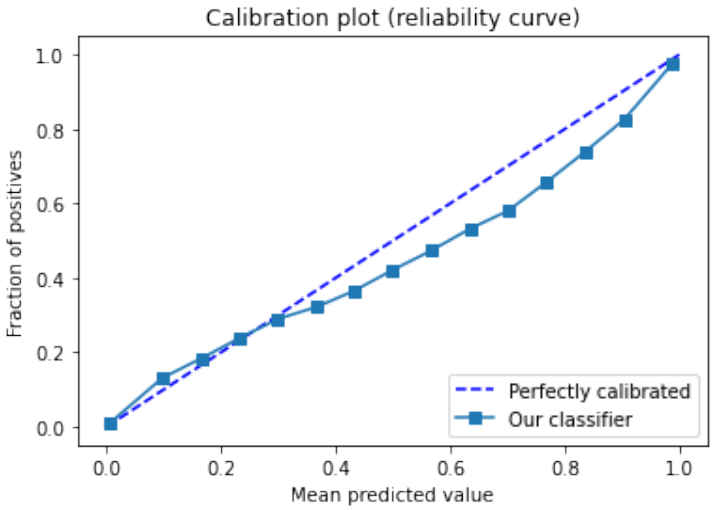

Figure 13: Calibration plot for the selected classifier.

to interpret and represent, so instead the predictions are aggregated in squares described in the Data Section. The densities (fraction of the area of each square with the label 'cocoa') of farms for each year are calculated. Figure 14 shows the estimated density of farms in 2013 and 2021, the background is NDVI band imagery of the region of interest.

Although some changes are visible, there is no clear pattern and also not all available imagery is used. We aggregate all available information by running the following regression for each square:

$$y_t = \alpha + \beta(t - 2013), \ t = 2013...2022$$

Where y_t is the predicted density in period t. This regression estimates a possible trend in the farms' density. If the coefficient β is positive it means the number of farms was growing over time and if β was negative, the number of farms decreased.

Out of 775 squares that are inside the area of interest, we select squares satisfying two criteria

1. average cocoa density $> 0.2\%$
2. the slope coefficient $\hat{\beta}$ is significant at the 5% level.

The first requirement ensures that we only analyse a change which is of substantial size (otherwise if the number of farms changed from 0 to 1 we will probably get significant results). The second criterion ensures that the change we observe is statistically significant. For all other squares, we assume a slope value of 0. See Section 4.3 for final maps and discussions.

4 Results

In Section 4.1 performance of the classifier is discussed, the variable importance graph is demonstrated and our estimates of annual crop area together with annual production are presented. Section 4.2 describes the simulation experiment validating our research. Section 4.3 presents

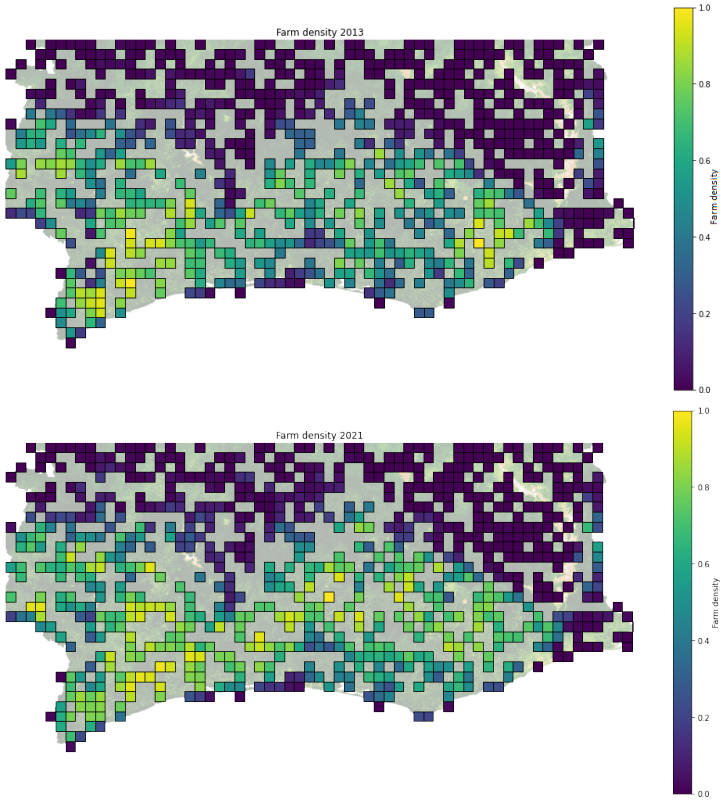

Figure 14: Farms location density in 2013 (top) and 2021 (bottom).

cocoa change maps and provides a discussion of obtained results. In Section 4.4 we discuss the impact of the EU deforestation law on cocoa production and the well-being of farmers.

4.1 Estimation of annual cocoa production and feature importance

We estimate farmland area, multiply it with an estimated crop yield and get a production estimate. We recommend using at least 32 GB of RAM to predict the pixels' labels. The feature importance is extracted to understand when the cocoa farms can be distinguished easier.

We evaluate the performance of a new classifier with a new threshold for a subsample not used in accuracy analysis. We achieve an accuracy of 94.5%, UA of 82.0% and PA of 85.8%. Figure 15 shows the variable's importance graphs for the selected XGBoost classifier. We see that the top 3 features are the same as in Figure 10 suggesting that our results are robust.

Below is Figure 16 with our final results, it includes the predicted area and production as well as production data from two different sources. All data is scaled to the area prediction in 2013. Overall, we see somewhat similar behaviour between the predicted production and the actual production. The correlation between the predicted production and FAOSTAT data is 0.618 and between the predicted production and Statista data is 0.621. The correlation between the two production datasets is 0.770, so our correlation coefficient seems to be reasonably high. It is worth mentioning that the correlation will be significantly higher if we remove the forecast for 2022, and in 2022 the pixel coordinates for the last two months are different from the previous month, so we replaced the values in November and December with values in October. At the start of 2023, the Ghana Cocoa Board (COCOBOD) said it expected the production to increase

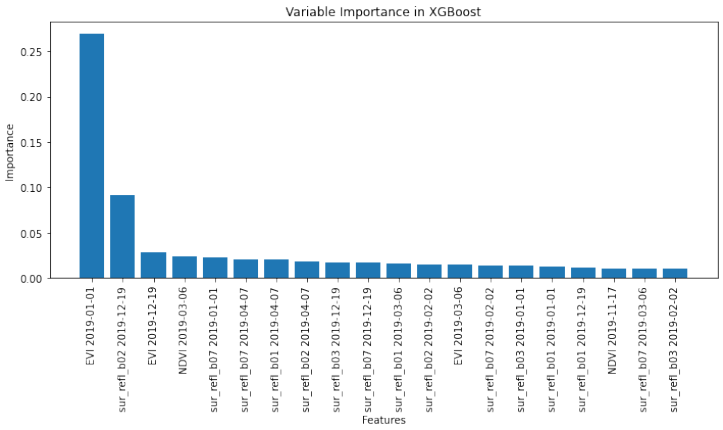

Figure 15: Variable importance graph XGBoost.

by 76% which might support our predicted increase in production in 2022 (Cocoa Board of Ghana, 2023).

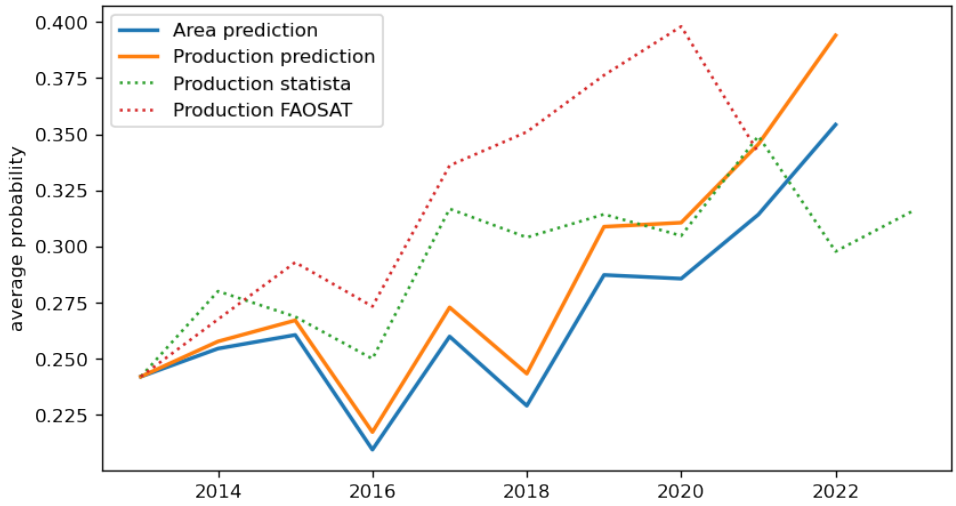

Figure 16: Average probability of crop area and production, all data is scaled to the first crop area probability value.

The most important band is by far the EVI band in January followed by the b02 band in December and EVI band in December. This result suggests that data in December and January is most helpful for cocoa separation. Production predictions can also be useful for further analysis involving, social, macroeconomic or climate change issues.

4.2 Prediction quality

A simulation experiment is implemented to see whether the obtained correlated values are statistically significant. The distribution of correlation values is analysed for the case when we are certain there is no correlation. This allows us to see the likelihood of the original correlation value under the assumption of no correlation. For this, a time series of prediction and production data is bootstrapped (so sampled with replacement) and the correlation coefficient between the artificial series is calculated. For example if the original series were (y_1, y_2, y_3, y_4) and $(\hat{y_1}, \hat{y_2}, \hat{y_3}, \hat{y_4})$,

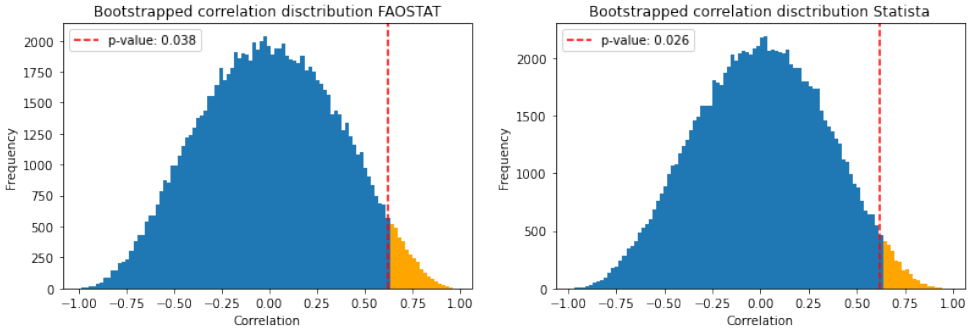

Figure 17: Boostrapped correlation values distribution, FAOSTAT (left) and Statista (right).

after bootsrapping a correlation coefficient between (y_3, y_2, y_1, y_4) and $(\hat{y_4}, \hat{y_1}, \hat{y_4}, \hat{y_2})$ might be recorded. This process is repeated 100,000 times, meaning production and prediction datasets consisting of 10 points each were bootstrapped and their correlation was computed 100,000 times. Given that after random sampling no correlation can be left, a distribution under the H_0 is: *no correlation between the original series*. The p-values of the observed correlation with H_a: *positive correlation is present* are calculated. Figure 17 summarises this experiment. We get p-values of 0.026 and 0.038 for Statista and FAOSTAT data respectively suggesting that our results are significant. This numerical experiment validates our results against available data.

4.3 Cocoa change maps

We analyse the change in cocoa farms' locations and build a theory that it is linked to climate change. Map 21 in Appendix D might be helpful to understand the geography of the regions we discuss.

Figure 18 shows the slopes of farms' change regression results. A pixel of non-purple colour means that the slope satisfied conditions in Section 3.7. Out of 775 squares, 169 satisfy these criteria. The top graph shows the scaled magnitude of slopes, meaning that the change in pixels labelled green and yellow was the highest (measured as an absolute increase in the farms' density per year). The graph below indicates any pixels where the conditions in Section3.7 were satisfied.

Out of 169 selected slopes, only one is negative, suggesting that over the 10-year period, the quantity of farms does not reduce in selected regions. This is the first insight of our analysis. One potential explanation could be that cocoa farmers actually do not abandon the areas if the unfavourable weather conditions decrease crops, rather farmers expand to the new areas while still keeping production in the original areas. This hypothesis contradicts with Yao Sadaiou Sabas et al. (2020) assumption that some territories were abandoned (although we should acknowledge that the authors run their analysis for a wider time range 1985-2019). The other more pragmatic explanation is that our classifier is unable to distinguish between a functioning cocoa farm and an abandoned cocoa farm.

Ruf et al. (2015) argue that between 1970 and 2000 an east-to-west shift is observable in Ghana and Ivory Coast farms, so most of the farms are located in the Central and Western parts of both countries. The Southwest of Ivory Coast has a more humid climate which is more favourable for cocoa beans and the prospect of higher rainfall is the main driver for cocoa

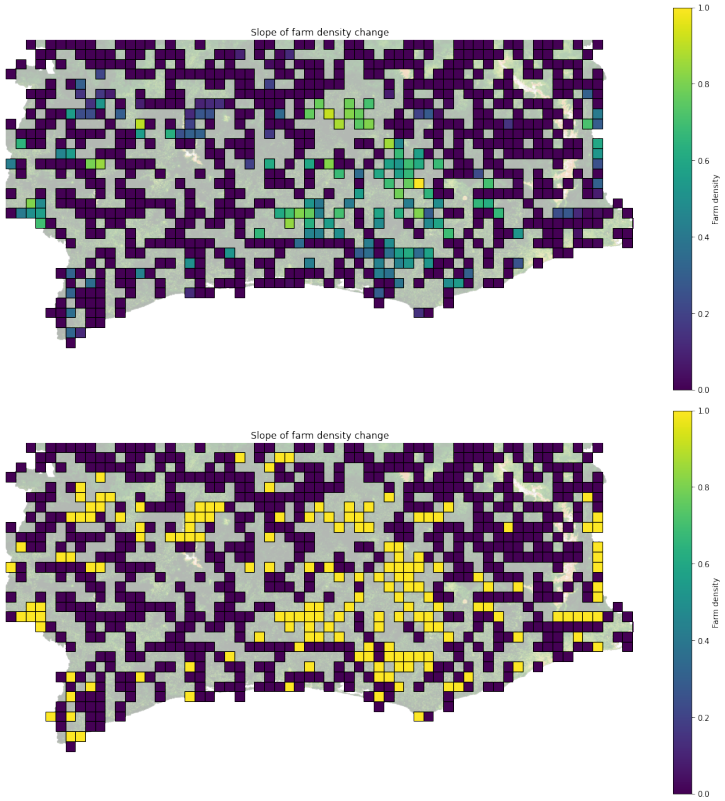

Figure 18: Farms migration map. Change in the density of farms, scaled (top), binary (bottom).

farmers changing location. This discussed pattern is generally confirmed by our located cocoa farms shown in Figure 14.

From the change of density maps in Figure 18 we see a spatial dependency between the points (so that groups of changed and not changed squares are grouped together). We observe some expansion in regions with high farm density. However, the majority of farms appeared on the border between West of Ghana and East of Ivory Coast. This is surprising given that these regions are considered rather unfavourable by Ruf et al. (2015). One possible explanation can be that in recent years, this region became more attractive due to climate change. The average temperature was rising for the Ivory Coast (World Bank, 2023) which is not particularly bad for cocoa, but warmer temperatures together with high humidity increase the spread of various diseases such as black pod (Asitoakor et al., 2022). Another risk is flood, when the cocoa roots are underwater they cannot breathe and it can cause them to rot. This is particularly relevant to southwestern regions such as Soubre and Agboville. Limited data is available to access this conjecture, but (Cilas, Goebel, Babin & Avelino, 2016) argue that such climate change might be heavily disturbing for agriculture in the tropics and might affect entire industries and several news articles on flooding in these regions can be found (see Reuters (2023) for example).

We believe that the observed migration is a cause of environmental change which makes planting cocoa in the West of Ivory Coast and Ghana more risky. Farmers are afraid to lose their crops due to diseases or floods and they move to the East of Ivory Coast and West of Ghana for a bit colder and less humid climate.

4.4 Cocoa production impact and EU deforestation law

In this section, we discuss the impact of the cocoa industry on the countries in Western Africa, illegal cocoa farms and the mechanism behind CO_2 emissions of cocoa farms and the implications of EU deforestation law.

When Ivory Coast was granted independence in 1960, it was covered in a jungle with hundreds of thousands of elephants, and an extensive population of chimpanzees along with many other species. Currently, it is estimated only 200-400 elephants are left, chimpanzees are considered endangered and only 4% of the country's land cover remains densely forested due to the entire country adapting to the cocoa industry (Higonnet et al., 2017). Many national parks and wildlife reserves exist with the purpose of preserving the environment in Ghana and the Ivory Coast.

In 2017, the environmental organisation Mighty Earth published a report showing that chocolate used by companies like Mars or Lindt was grown illegally in national parks (Higonnet et al., 2017). It is estimated that more than 90% of the land mass inside protected areas is covered by cocoa, meaning that areas where cocoa plantations are prohibited, ironically, consist mostly of cocoa. Mighty Earth found evidence that large EU cocoa traders such as Cargill and Olam purchase this illegally grown cocoa (Higonnet et al., 2017).

Cocoa is responsible for a significant share of tropical deforestation and CO_2 emissions (forests soak up the carbon dioxide from the atmosphere and when the trees are cut and processed they release CO_2 back into the atmosphere). Deforestation contributes to the increase in CO_2 levels and impacts climate change. Cocoa leads to deforestation in two possible ways. First, the increasing global demand for cocoa incentivises the farmers to expand (and the way to do it is to cut tropical forests and plant cocoa). Our research estimates an increase in the production area almost every year between 2013 and 2022. However, there is one more effect that can be attributed to the deforestation. Due to climate change, the farmers might have to move around the country to find regions with more favourable climate conditions. This creates a vicious circle - an increase in the temperature leads to less cocoa produced leading to less income for the farmers. Because of that, farmers are forced to expand, meaning that the new forest is cut and the old regions where cocoa was previously planted are abandoned (Cocoa Life, 2021). This deforestation further contributes to the CO_2 emissions and greenhouse effect leading to an increase in temperature and forcing farmers to move again in a few years. Although we did not find evidence that farms are being abandoned, it is known that the crop yield of some regions is affected by climate change (Reuters, 2023), effectively forcing farms to expand by the same mechanism as described.

Apart from climate change, the working conditions at these farms are not good as benefits from cocoa production are skewed towards cocoa traders and redistributors (Cocoa Life, 2021). The majority of farm workers are paid less than 1$ a day which is far below the poverty line. There are also many documented cases of slavery and child labour at cocoa farms. Food Empowerment Project (2022) estimates that 2.1 million of children work at cocoa farms in Ghana and Côte d'Ivoire. These children often do not get basic education and are held in poor conditions. The authors argue that child labour might be the only way for farms to keep prices competitive. Remote sensing can help to locate areas with potential child labour exploitation. For example, an expanding farm in a protected area might experience a shortage of workforce

and child labour exploitation could follow. Our research provides a framework for such analysis.

On the 16th of May 2023, governments across the EU adopted a law that some commodities linked to deforestation or forest degradation cannot enter the EU after 2024. These commodities include palm oil, cattle, soy, coffee, timber, rubber and cocoa. Some of the largest EU producers import cocoa planted illegally in the national parks and the EU values cocoa imports at multiple billion dollars. We see that the production of cocoa is associated with large externalities influencing CO_2 emissions, endangered species and child labour.

Cancelling cocoa imports will have a substantial negative impact on the affected population's well-being. With the EU restricting the main source of income for the affected populations, it is essential to provide sufficient support for local populations in order to ensure their transition to alternative sources of income. In 2021 €18 million were committed to Côte d'Ivoire as a part of the "Sustainable Cocoa" program. In Ghana, the EU contributions amounted to €12 million up to 2023 (European Commission, 2022). However, these amounts might not be sufficient to leave a significant impact on millions of workers who will be losing their main source of income. In October 2022, the EU pledged to raise €450 million to fight child labour and deforestation in the Ivory Coast. Although this is a great initiative, it is essential to ensure these funds will reach those in need. It is challenging to locate cocoa farms, especially illegal farms and satellite imagery might help to locate the regions where this aid is needed the most.

5 Conclusion

Due to the lack of available data, small size and similarity to forest cocoa farms are not easy to detect. Yet, extensive knowledge about cocoa farms' locations might help to accurately estimate total cocoa production, including in illegal areas. It can help assess deforestation and CO_2 emissions caused by cocoa farms and estimate changes in cocoa farms' landscape. Locating cocoa farms also allows for finding and supporting people working on these farms.

Using a cocoa map created by (Kalischek et al., 2022) we have created a classifier that performs in line with or even outperforms the existing cocoa farms classifiers. We succeed due to the use of non-conventional vegetation imagery, stacking the images in different months together without 'shifting' and of course higher quality of labelled data.

By tracking the number of farms over time we are able to see trends in cocoa farms' crop area and production and via the simulation experiment, a significant correlation with the available production data is confirmed. We are also able to extract the most important features and conclude that the months of January and December are the most helpful in identifying cocoa plantations.

We recreate cocoa maps and the cocoa locations agree with the existing research of Abu et al. (2021); Kalischek et al. (2022) and Ruf et al. (2015). We track the change in the location of cocoa farms. First, we find no evidence that in some regions cocoa farms are abandoned completely, as we almost never observe a significant decrease in the number of farms in selected regions. Secondly, we observe the appearance of new farms towards the East of the Ivory Coast and West of Ghana. These regions are generally considered less favourable for cocoa plantations. However, we believe that some farmers prefer to expand in such areas as they are less exposed to the risks of flood and cocoa diseases. Furthermore, changes in the location of cocoa farms

might go together with changes in workforce demography and even changes in child labour and slavery. Hence, various social insights can be gained from such analysis.

Côte d'Ivoire and Ghana used to have a beautiful nature with a variety of species and these countries suffered substantially in the last century from the cocoa industry. The benefits of chocolate production are concentrated around EU and US companies and distributors, while in Western Africa millions of people are involved in the cocoa industry and living below the poverty line. Some workers on cocoa farms have never tasted chocolate. There are entire cocoa villages with schools and churches located in the national parks harvesting illegal cocoa (Higonnet et al., 2017).

We believe the EU deforestation law will help to reduce CO_2 emissions, increase biodiversity and help fight child labour and poverty. However, we should remember that with this law, a part of the cocoa business involving millions of people will become illegal. These people did not choose to work on these illegal farms and they should be given plenty of support and opportunities to transition into new businesses. To detect where the aid must be provided it is important to be aware of the cocoa farms' locations and due to the lack of other sources, satellite imagery appears to be the only near-real-time way to locate the farms.

6 Limitations and future research suggestions

The first restriction we have faced is computing limitations. Although this research was performed on relatively advanced cloud servers, having better computing power will allow us to gain deeper insights. Sentinel imagery available in higher resolution can be used instead of Modis imagery. This will potentially allow for higher precision and more localised inference. With better memory, it is also possible to create complete cocoa maps. Finally, more compute power will allow an increase in the amount of data points used (the presented classifiers were trained on just below 2,000,000 randomly sampled data points). Without these limitations, the scale of this research will be comparable to the one of Khachiyan et al. (2021). For future research, the analysis scale might be increased and creating more precise and complete cocoa maps over time can be a good starting point.

We have trained the classifier using predicted labels from another research paper. Hence, our classifier is as good as the map from Kalischek et al. (2022). Moreover, this approach can potentially inflate the accuracy results as the most 'difficult cases' might be misclassified by the original paper. Our research might also suffer from spatial autocorrelation - an effect where overfitting might occur as train and test points located next to each other exhibit higher similarity than more distant points. We have validated the performance of our classifier on the out-of-sample parts of the same model, but we do not have tools to evaluate the accuracy across time (except for validating the predicted crop area).

Similarly, it is difficult to directly validate the graph with the farms' location change (Figure 18). One might suspect that it is not insightful if the predicted farms' locations are similar to the farms' locations in the training set. Yet, this result actually validates our research. Our algorithm is not 'aware' of the location of the input data and only a small fraction of data is used for training, meaning that even for 2019 (the year when we train our classifier) a map was mostly constructed out of sample.

The next logical step can be establishing a better connection between the cocoa change map and child labour. It might be possible to find a data-driven link between the cocoa farm changes and documented child labour cases which can be the first step in tackling these issues. Similarly, satellite images will be helpful in observing the consequences of the EU deforestation law by creating a cocoa change map in protected areas. By tracing the change in cocoa production in 2023 and 2024 it might be possible to estimate the true impact and unemployment numbers caused by this law.

Finally, we would like to clarify that our explanation about cocoa farms migration caused by climate change, flooding and cocoa diseases is a conjecture, there is no direct evidence supporting this claim. Combining our results with qualitative insight based on surveys, interviewing or industry expertise might result in more precise explanations and serve as a basis for policy recommendations.

7 Code description

The code is written in Python and it requires a lot of interaction with data files. Since the data files take hundreds of gigabytes of space, they are thus excluded from research submission. However, all code is self-contained; by running files in the correct order, all relevant output will be saved in the correct folders. The final classifier is not attached as well because it occupies 119 MB. We would, however, be happy to supply it and some data files upon request. Furthermore, running the code requires a developer account of Google Earth Engine API.

All code is written in Jupyter Notebooks. Every notebook is responsible for a part of the research and should be run in the following order. Firstly, the *Training data collection.ipynb* collects and transforms 1,000,000 data points of both 'cocoa' and 'not cocoa'. Secondly, the *Model selection.ipynb* contains cross-validation, best classifier analysis and threshold analysis. The next file - *Testing data collection and results.ipynb*, shows time series data collection, bootstrap simulation and final predictions graphs. *Cocoa migration and climate change.ipynb* contains cocoa maps and cocoa change maps as well as regression for every square. Finally, *Pretty images.ipynb* presents some illustrative analyses that are not directly relevant to this research, but were used in the process. It also includes a linear trend regression on the cocoa crop yield.

References

Abu, I.-O., Szantoi, Z., Brink, A., Robuchon, M. & Thiel, M. (2021). Detecting cocoa plantations in Côte d'Ivoire and Ghana and their implications on protected areas. *Ecological Indicators*, *129*, 107863. doi: https://doi.org/10.1016/j.ecolind.2021.107863

Asitoakor, B. K., Asare, R., Ræbild, A., Ravn, H. P., Eziah, V. Y., Owusu, K., ... Vaast, P. (2022). Influences of climate variability on cocoa health and productivity in agroforestry systems in Ghana. *Agricultural and Forest Meteorology*, *327*, 109199. doi: https://doi.org/10.1016/j.agrformet.2022.109199

Batista, J. E., Rodrigues, N. M., Cabral, A. I. R., Vasconcelos, M. J. P., Venturieri, A., Silva, L. G. T. & Silva, S. (2022). Optical time series for the separation of land cover types with similar spectral signatures: cocoa agroforest and forest. *International Journal of Remote Sensing*, *43*(9), 3298-3319. doi: https://doi.org/10.1080/01431161.2022.2089540

Bergstra, J. & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, *13*(2). `http://www.jmlr.org/papers/v13/bergstra12a.html`.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*, 123–140. doi: https://doi.org/10.1007/BF00058655

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. doi: https://doi.org/10.1023/A:1010933404324

Centre for the Promotion of Imports from Developing Countries. (2022). *What Demand Exists for Cocoa?* `https://www.cbi.eu/market-information/cocoa/what-demand`.

Chen, T. & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Cilas, C., Goebel, F.-R., Babin, R. & Avelino, J. (2016). Tropical Crop Pests and Diseases in a Climate Change Setting—A Few Examples. In E. Torquebiau (Ed.), *Climate change and agriculture worldwide* (pp. 73–82). Dordrecht: Springer Netherlands. doi: https://doi.org/10.1007/978-94-017-7462-8_6

Cocoa Board of Ghana. (2023). Retrieved: June 17, 2023, from `https://cocobod.gh/`.

Cocoa Life. (2021). *Cocoa Life: Climate Change.* `https://www.cocoalife.org/the-program/climate-change/`.

Douzas, G., Bacao, F., Fonseca, J. & Khudinyan, M. (2019). Imbalanced learning in land cover classification: Improving minority classes' prediction accuracy using the geometric SMOTE algorithm. *Remote Sensing*, *11*(24), 3040. doi: https://doi.org/10.3390/rs11243040

European Comission. (2023). *Cocoa map for Côte d'Ivoire and Ghana.* `https://africa-knowledge-platform.ec.europa.eu/explore_maps?title=Cocoa0map0for0Cote0d7Ivoire0and0Ghana`.

European Commission. (2022). *EU, Côte d'Ivoire, Ghana, and Cocoa Sector Endorse Alliance for Sustainable Cocoa.* `https://policy.trade.ec.europa.eu/news/eu-cote-divoire-ghana-and-cocoa-sector-endorse-alliance-sustainable-cocoa-2022-06-28_en`.

Filella, G. B. (2018). Cocoa segmentation in Satellite images with deep learning. `https://ethz.ch/content/dam/ethz/special-interest/baug/igp/photogrammetry-remote`

-sensing-dam/documents/pdf/Student_Theses/BA_BonetFilella.pdf.

Food and Agriculture Organization of the United Nations. (2020). *The State of the World's Forests.* https://www.fao.org/state-of-forests/en/#:~:text=Between%202015%20and%202020%2C%20the,80%20million%20hectares%20since%201990.

Food and Agriculture Organization of the United Nations. (2023). *FAOSTAT: Food and Agriculture Data.* Retrieved: May 15, 2023, from https://www.fao.org/faostat/en/#data/QCL.

Food Empowerment Project. (2022). *Slavery in the Chocolate Industry.* https://foodispower.org/human-labor-slavery/slavery-chocolate/.

Goldbloom, A. (2016). *What Algorithms are Most Successful on Kaggle?* https://www.kaggle.com/code/antgoldbloom/what-algorithms-are-most-successful-on-kaggle.

Google Earth Engine Team. (2023). *MODIS collection 6.* Retrieved: May 15, 2023, from https://developers.google.com/earth-engine/datasets/catalog/MODIS_061_MOD13A1.

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning.* New York, NY, USA: Springer New York Inc.

Higonnet, E., Bellantonio, M. & Hurowitz, G. (2017). *Chocolate's Dark Secret.* https://policycommons.net/artifacts/3446820/chocolates_dark_secret_english_web/4246954/. United States of America.

Kakaoplattform. (2022). *Cocoa facts and figures.* Retrieved: June 10, 2023, from https://www.kakaoplattform.ch/about-cocoa/cocoa-facts-and-figures.

Kalischek, N., Lang, N., Renier, C., Daudt, R. C., Addoah, T., Thompson, W., ... Wegner, J. D. (2022). Satellite-based high-resolution maps of cocoa for Côte d'Ivoire and Ghana. *Nature Food*. doi: https://doi.org/10.48550/arXiv.2206.06119

Khachiyan, A., Thomas, A., Zhou, H., Hanson, G. H., Cloninger, A., Rosing, T. & Khandelwal, A. (2021, December). *Using Neural Networks to Predict Micro-Spatial Economic Growth* (Working Paper No. 29569). National Bureau of Economic Research. doi: https://doi.org/10.3386/w29569

McElfresh, D., Khandagale, S., Valverde, J., C, V. P., Ramakrishnan, G., Goldblum, M. & White, C. (2023). *When Do Neural Nets Outperform Boosted Trees on Tabular Data?* doi: https://doi.org/10.48550/arXiv.2305.02997

Pendrill, F., Persson, U. M., Godar, J., Kastner, T., Moran, D., Schmidt, S. & Wood, R. (2019). Agricultural and forestry trade drives large share of tropical deforestation emissions. *Global environmental change*, *56*, 1–10. doi: https://doi.org/10.1016/j.gloenvcha.2019.03.002

Quadros, F. D., Snellen, M., Sun, J. & Dedoussi, I. C. (2022). Global civil aviation emissions estimates for 2017–2020 using ADS-B data. *Journal of Aircraft*, *59*(6), 1394–1405. doi: https://doi.org/10.2514/1.C036763

Radio2Space. (2013). *Components of Electromagnetic Spectrum.* https://www.radio2space.com/components-of-electromagnetic-spectrum.

Reuters. (2023). Heavy Rain in Ivory Coast Raises Cocoa Disease Fears, Farmers Say. *Reuters*. https://www.reuters.com/markets/commodities/heavy-rain-ivory-coast-raises-cocoa-disease-fears-farmers-say-2023-06-05/.

Ritchie, H. & Roser, M. (2021). Forests and Deforestation. *Our World in Data*. https://ourworldindata.org/forests-and-deforestation.

Ruf, F., Schroth, G. & Doffangui, K. (2015). Climate change, cocoa migrations and deforestation in West Africa: What does the past tell us about the future? *Sustainability Science*, *10*, 101–111. doi: https://doi.org/10.1007/s11625-014-0282-4

Scheffers, B. R., Joppa, L. N., Pimm, S. L. & Laurance, W. F. (2012). What we know and don't know about Earth's missing biodiversity. *Trends in Ecology Evolution*, *27*(9), 501-510. doi: https://doi.org/10.1016/j.tree.2012.05.008

Statista. (2023). *Production of Cocoa Beans in Ghana.* Retrieved: June 2, 2023, from https://www.statista.com/statistics/497844/production-of-cocoa-beans-in-ghana/.

Wessel, M. & Quist-Wessel, P. F. (2015). Cocoa production in West Africa, a review and analysis of recent developments. *NJAS - Wageningen Journal of Life Sciences*, *74-75*, 1-7. doi: https://doi.org/10.1016/j.njas.2015.09.001

Wolfe, P. (1969). Convergence Conditions for Ascent Methods. *SIAM Review*, *11*(2), 226–235. doi: https://doi.org/10.1137/1011036

World Bank. (2023). *Climate Knowledge Portal - Côte d'Ivoire Climate Data (Historical).* Retrieved: June 1, 2023, from https://climateknowledgeportal.worldbank.org/country/cote-divoire/climate-data-historical.

Wu, J. (2017). Introduction to convolutional neural networks. *National Key Lab for Novel Software Technology. Nanjing University. China*, *5*(23), 495.

Yao Sadaiou Sabas, B., Gislain Danmo, K., Akoua Tamia Madeleine, K. & Bogaert, J. (2020). Cocoa Production and Forest Dynamics in Ivory Coast from 1985 to 2019. *Land*, *9*(12). doi: https://doi.org/10.3390/land9120524

A Used imagery examples

Below you can see an example of satellite images that are used for analysis. These images cover around 500 000 km^2 and are originally provided in .TIF format. Due to the high memory load (about 0.5 GB per image), they are converted to .PNG format. The size of the images is 9.47^o x 4.79^o which is 1052 x 532 km. This region contains all areas suitable for cocoa beans production in the countries of interest (Kalischek et al., 2022). In our analysis we do not operate on a downloaded image, rather we have images as abstract objects in the Earth engine API and only retrieve the relevant features.

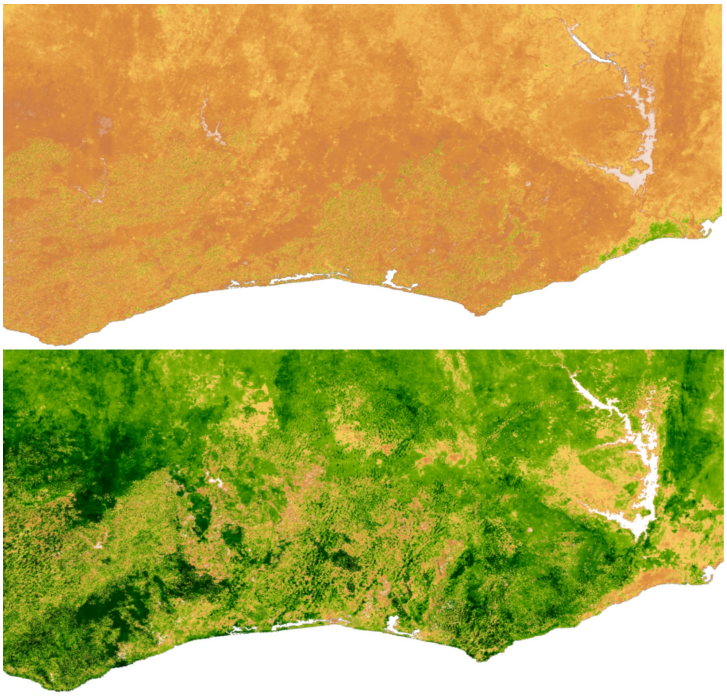

Figure 19: Imagery of Ghana and Ivory Coast coastline, NDVI and b07 bands.

B The meaning of colour in this research

Typically an image is considered a 3-dimensional RGB array where every pixel of an image holds values between 0 and 256 representing a corresponding colour intensity. However, in the physical world, these dimensions do not exist. What we see is a reflection of light of a certain frequency that our eyes convert to colour. Yet, human eyes can only capture a limited wavelength as can be seen in Figure 20.

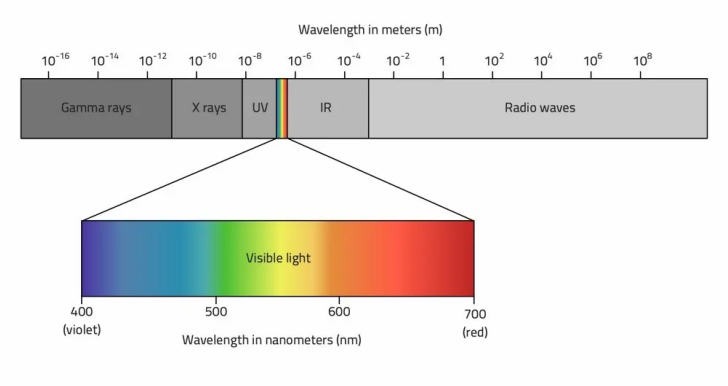

Figure 20: Electromangetic spectrum, retrieved from Radio2Space (2013).

Using spectral signature instead of RGB imagery not only allows the capturing of a wavelength not visible to a human eye but also allows a significant dimension reduction as now every band is mapped to a single number.

C Gradient descent

Here, we would like to briefly outline the idea of the gradient descent method.

Consider a convex twice differentiable function $f(x), \mathrm{f} : \mathbb{R}^n \to \mathbb{R}$, we are interested in solving $f'(x) = 0$. Assume we start with the initial point x_0 and consider a second-order Taylor expansion around this point.

$$f(x_0 + s) = f(x_0) + \nabla f(x_0)^T s + \frac{1}{2} s^T H_f s + \mathcal{O}(\|s\|^3) \tag{4}$$

So that we can say $f(x_0 + s) \approx f(x_0) + \nabla f(x_0)^T s + \frac{1}{2} s^T H_f s$. If we solve the minimisation problem for s we will arrive at Newton's Method. Yet, Newton's algorithm is more computationally expensive as it requires inverting the Hessian and that's where the Gradient Descent algorithm is helpful.

In gradient descent, we repeatedly use a first-order expansion only: $f(x_0 + s) \approx f(x_0) + \nabla f(x_0)^T s$. If we solve the minimisation problem with respect to s we would arrive at $-\infty$ (because it is a linear function in s, by construction). Gradient descent allows us to avoid it by restricting the values of s as $\|s\| \leq const$. In this case, the optimal solution for s will take the form of $-\alpha \nabla f(x_0)$. The $-\nabla f(x_0)$ part comes from the optimal direction of s as by minimizing $\nabla f(x_0)^T s$ with the constrained length of s, we solve $\nabla f(x_0)^T s = \|\nabla f(x_0)\| \|s\| cos(\nabla f(x_0), s)$ and in order to obtain $cos(\nabla f(x_0), s) = -1$ we require the direction of s to be the opposite of $\nabla f(x_0)$. The α can then be interpreted as the step size, or the neighbourhood size where we use the approximation $f(x_0 + s) \approx f(x_0) + \nabla f(x_0)^T s$.

The parameter α is called the learning rate. Under rather unrestrictive Wolfe conditions (Wolfe, 1969) the convergence of the algorithm outlined above is guaranteed after a certain number of steps. In practice, of course, many complications might arise (non-convex problems with multiple peaks, slow convergence, numerical issues).

D Map with borders

Figure 21: Map of Ghana and Ivory Coast with NDVI band background.