Cropland Mapping with satellite imagery and machine learning

Muhamed Tuo
tuomuhamed@gmail.com

Abstract. Crop maps play a pivotal role in agricultural monitoring and food resource management, while also offering valuable assistance for domain-specific application in emerging nations. By harnessing the potential of machine learning (ML) models and readily accessible satellite imagery, it becomes feasible to generate economically efficient and highly detailed crop maps. This aims to present a methodology to solving Cropland Mapping at 10m resolution, by using free-accessible satellite data from Sentinel ½ and Landsat time series data.

1. Introduction

Timely and precise crop maps are vital for applications in agriculture, as well as related fields like natural resources, environment, health, and sustainability. These maps are fundamental for practical agricultural use. Several algorithms and freely available products offer global cropland extent maps at a 30m resolution. However, these data have limitations, including infrequent updates, differing definitions of "cropland," and notable discrepancies among existing global cropland masks.

With new earth observation plans underway, more high-resolution satellite imagery is becoming available, and machine learning and artificial intelligence hold the potential to enhance land cover classification accuracy using this data.

In the effort of addressing these challenges and advancing global high-resolution cropland mapping through remote sensing, this initiative seeks to create precise and cost-effective classification models for cropland extent mapping using machine learning techniques. Participation in this initiative allows researchers and practitioners to contribute to the progress of global cropland mapping, facilitating a more accurate and comprehensive understanding of agricultural landscapes worldwide.

2. Study area

For this study, we focus on 3 regions of interest, namely the Nangarhar province (Afghanistan), the Khouzistan province (Iran) and the Al Hasahisa province (Sudan). The temporal cropland extent distribution for these region ate the following:

- Afghanistan: April 2021 ~ April 2022
- Iran : July 2019 ~ June 2020
- Sudan: July 2019 ~ June 2020

3. Dataset

We used the “Harmonized Sentinel-2 MSI: MultiSpectral Instrument, Level-2A” dataset, which is accessible on Google Earth through the following link .

4. Methodology

4.1. Data acquisition

We were given 500 data points for each region, and we utilized them to gather the corresponding Sentinel-2 time series. After filtering the data to match the specified time frame and ensure a minimum of 10% cloud-free coverage, we proceeded to select the 12 "B" Bands (B1 to B12), the 3 True Color bands (TCI_R, TCI_G, TCI_B), and 2 additional noteworthy bands (AOT and WVP).

Bands name	**Wavelength**	**Description**
B1		Aerosols
B2		Blue
B3		Green
B4		Red
B5		Red Edge 1
B6		Red Edge 2
B7		Red Edge 3
B8		NIR
B8A		Red Edge 4
B9		Water Vapor
B11		SWIR 1
B12		SWIR 2
AOT		Aerosol Optical Thickness
WVP		Water Vapor Pressure

Table 1: Sentinel-2 bands and description

4.2. Feature engineering

We employed vegetation indices as a method for generating features to serve as inputs for the model. Using the 12 "B" bands, we computed 50 vegetation indices, including NDVI, EVI, and SAVI. Given that the data constitutes a time series, we had N observations for each location. To aggregate these N observations, we initially applied three descriptive statistics functions (mean, max, min). Subsequently, we opted to utilize only the "min" function as an aggregator because it exhibited superior performance on its own compared to the other two methods. As a result, the final input for the model comprises the initial 17 raw bands, along with the 50 computed vegetation indices. We then splitted the data by region and discarded the location information in order to ensure the robustness and generalizability of the model.

4.3. Modeling

Given the problem definition, the inherent structure of both the initial and final/processed data, we opted for the choice of Gradient Boosting Algorithm.

We trained 3 different Boosting Algorithms (Catboost, LightGBM, XGboost) for each of the regions and evaluated the performance difference between them and also to estimate the complexity of the task to solve. All the 3 models showed a great result for Iran and Sudan, while underperforming on the Afghanistan data. We continued our experiments with Catboost as it was the best model on average.

To ensure the robustness of the final model, we followed a 5 K-fold strategy to train and choose the best experiment. Thus, the final model is an ensemble of the 5 fold models, each trained on a portion of the data. The final output is a simple weighted average of the 5 models outputs.

	Afghanistan	Iran	Sudan	Average
Validation	0.876	0.956	0.974	0.935
Public LB	X	X	X	0.94
Private LB	X	X	X	0.946

Table 2 : Performance of the final model on the validation and test sets, based on the accuracy metric.

5. Accuracy vs Time interval

In the case of both Iran and Sudan, the time interval spanning "July 2019 to June 2020" produced the best results, achieving an accuracy exceeding 95%. However, expanding this interval to include more historical data led to a decline in model performance on both the validation and test datasets.

Regarding Afghanistan, utilizing only the data available for April 2022 posed significant challenges, as the model could only achieve an accuracy of 86% for that specific time frame. Extending this period to encompass more data, from April 2021 to April 2022, resulted in an improvement of +2% on the validation set and approximately +0.7% on the test set.

6. Conclusion

In this research, we demonstrated the efficacy of cropland mapping using satellite imagery, achieved through the fusion of vegetation indices and machine learning algorithms. Remarkably, we achieved competitive and promising results despite having only 500 data points per region. It is worth noting that acquiring a larger and more precise dataset has the potential to yield substantial performance improvements.