Cropland mapping in KAZA using Earth Observation and Machine Learning

Alexander Merdian-Tarko, January 2023

The location and distribution of cropland as well as its evolution over time are key information for successful impact monitoring efforts in nature conservation, especially in areas such as the Kavango-Zambezi Transfrontier Conservation Area. Machine Learning algorithms and Earth Observation data can be leveraged to create annual crop maps efficiently at scale, indicating the distribution of cropland in a given region of interest. Fed with existing cropland locations (KAZA Land Cover 2020 dataset) and using publicly available satellite imagery (e. g. Sentinel-2), random forest and NASA Harvest's OpenMapFlow are able to reach the baseline classification accuracy of 75%. The resulting crop maps, however, show inconsistencies in the predicted cropland distribution and area - both in a comparison across models and also considering the used baseline land cover dataset. Using higher-quality training labels (e. g. new farm plot locations collected in the field) and higher-resolution satellite imagery (e. g. Planet) could potentially be a major step towards improving the quality of the produced crop maps. More accurate crop maps could then contribute to better supporting impact monitoring efforts on the ground.

Introduction and context

Knowing where cropland is located and how its location and extent change over time is critical for impact monitoring efforts to answer questions related to topics such as sustainable agriculture, deforestation, biodiversity and climate change. Machine Learning (ML) algorithms can help identify cropland on a large scale efficiently using Earth Observation (EO) data, more precisely in this context publicly available satellite imagery. A Machine Learning algorithm is trained on existing data of pixel-based cropland locations in satellite images. Common challenges include the availability of high-quality training labels (i. e. cropland or farm plot locations), the availability of high-resolution satellite imagery, the fact that the chosen study area exhibits mostly small-scale agriculture and the relatively common abandonment of farm plots to seek out new fertile land. The trained model can be used to create annual crop maps that show the amount and distribution of cropland in a given region. Crop maps of the same region for different years can be used to explore changes over time and uncover trends in the amount and distribution of cropland. These region-specific annual crop maps can inform better decision-making and thus support impact monitoring efforts on the ground.

Generating crop and non-crop points from existing data

The KAZA Land Cover 2020 dataset is used to generate a balanced sample of 1,000 crop and 1,000 non-crop points that function as the input labels for the applied ML models. The six Bengo regions within the Kavango-Zambezi Transfrontier Conservation Area (KAZA) are of particular interest. Therefore, the crop and non-crop points are sampled from the six

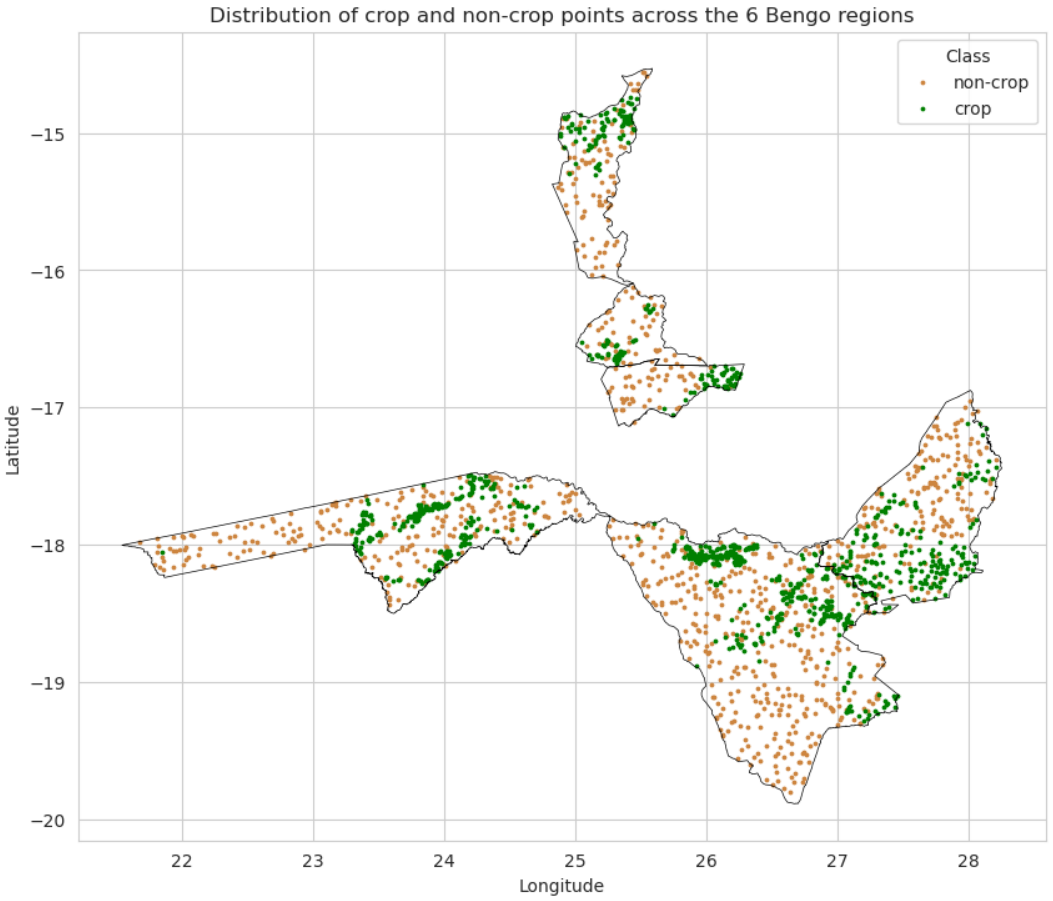
Bengo regions exclusively. The amount of cropland in the whole of KAZA is 4% whereas it is 9% in the six Bengo regions. The amount of cropland varies quite significantly from region to region (see **Table 1**). The total area per region is calculated using Google Earth Engine based on a set of provided shape files. The crop area share is derived from the relative amount of crop pixels in the KAZA Land Cover 2020 dataset (see this notebook for the detailed computations).

Table 1: Total area and crop area for the six Bengo regions according to the KAZA Land Cover 2020 dataset.

	Total area	Crop area
Binga	13,420.93 km²	3,498.84 km² (26.07%)
Hwange	27,395.88 km²	980.77 km² (3.58%)
Mufunta	6,438.26 km²	377.28 km² (5.86%)
Mulobesi	3,595.73 km²	36.67 km² (1.02%)
Sichifulo	3,558.19 km²	519.85 km² (14.61%)
Zambezi	17,113.67 km²	652.03 km² (3.81%)

The sampled crop class comprises solely the cropland land cover class. The accuracy of the cropland land cover class in the KAZA Land Cover 2020 dataset is 75% based on a hold-out set that contains 30% of the total data (more details can be found in the respective technical report). This is the baseline accuracy against which the applied models are compared. The non-crop class consists of all other 17 available land cover classes which are not uniformly distributed though. Some appear much more often than others; some don't appear in specific regions at all. The amount of points sampled per region is proportional to its area. The sampling procedure can be explored in more detail here. **Figure 1** shows the 1,000 crop and 1,000 non-crop points sampled randomly from the KAZA Land Cover 2020 dataset for the six Bengo regions.

Figure 1: 1,000 crop and 1,000 non-crop points sampled randomly from the KAZA Land Cover 2020 dataset for the six Bengo regions.



Models

Random forest is a well-established ensemble model based on individual decision trees (10 estimators in the present case) and widely used in different fields, amongst others land cover classification. The input data comprises a Sentinel-2 mean composite for the period of 2020-01-01 to 2020-12-31, using bands B2, B3, B4, B8 and NDVI. The input data covers only a period of 12 months due to computational limitations within Google Earth Engine. Random forest is trained on points that are aggregated at a yearly resolution (i. e. one time step).

OpenMapFlow is a Python package developed by NASA Harvest to rapidly create maps using ML and EO. It consists of a command line interface and Jupyter notebooks providing different functionalities such as processing data, training models and creating maps. It uses Google Earth Engine to obtain EO data and Google Cloud for storage and computational resources. Users can use existing datasets but can also create their own datasets by providing longitude, latitude and class labels of the target (e. g. cropland, maize or buildings) for a given point in time. In the present case a custom dataset is created using the sampled 1,000 crop and 1,000 non-crop points described above. Both existing

(GeowikiLandcover2017) and custom data (KAZABengoCrop2020Random2000) are jointly fed into a pre-trained PyTorch Deep Learning model that leverages the time series nature of the satellite input data. The features of the monthly satellite time series include several Sentinel-2 bands, several Sentinel-1 bands, precipitation and temperature from ERA5, elevation and topography from SRTM DEM and NDVI. The monthly time series consists of 24 time steps within the period from 2020-01-01 to 2021-12-31.

The main differences between the two applied models are the following:

- Random forest is an ensemble model based on decision trees whereas OpenMapFlow leverages Deep Learning.
- Random forest considers only one point in time whereas OpenMapFlow employs satellite data as a monthly time series.
- Random forest uses data from one satellite and five features whereas OpenMapFlow uses four data sources and 18 features.

Evaluation

The balanced validation and test sets used for evaluation and model comparison are created randomly within OpenMapFlow and consist of 378 and 428 observations, respectively. **Figure 2** shows validation and test set accuracies at different classification thresholds for both applied models. Accuracy is a suitable metric in the present case since a balanced dataset was purposely created. The default classification threshold of 0.5 appears to be a sensible choice in general as it yields (close to) peak accuracy. The performance of both models seems to be comparable in terms of predictive accuracy.

Figure 2: Validation and test set accuracies at different classification thresholds for random forest and the OpenMapFlow model.

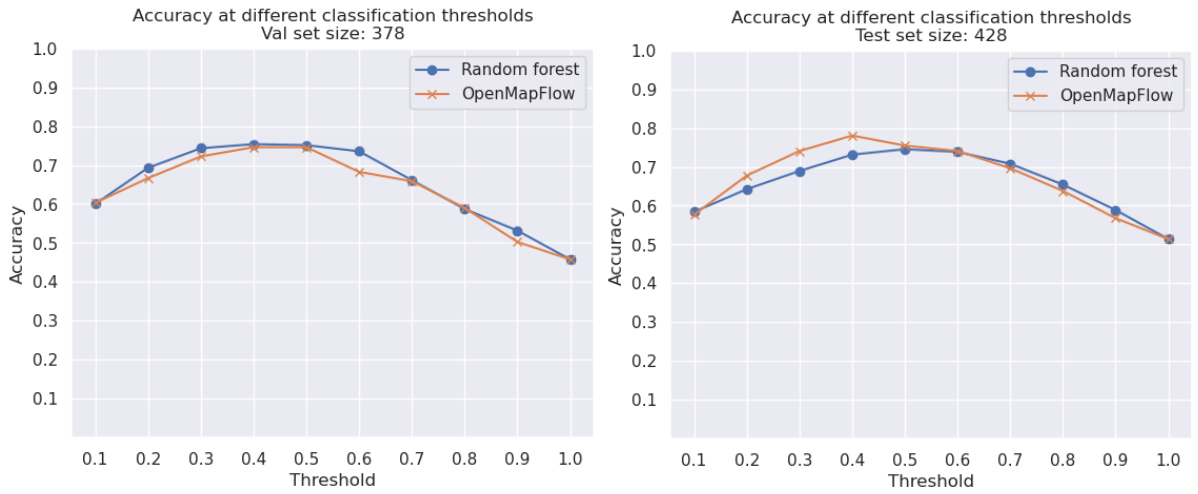


Table 2 shows additional classification metrics to paint a more complete picture. Both models show a similar performance in terms of all considered classification metrics. The model performance seems to be generally consistent across both the validation and the test

set although there are slight variations which are probably due to somewhat different data within both subsets. Both models reach the KAZA Land Cover 2020 dataset's cropland baseline accuracy of 75%. The fact that both models do not exceed the baseline accuracy reflects the importance of high-quality input labels. Improving label quality could thus potentially increase predictive accuracy in turn.

Table 2: Validation and test set performance metrics for random forest and the OpenMapFlow model (classification threshold: 0.5)

	Random forest		OpenMapFlow	
	Validation set	Test set	Validation set	Test set
Accuracy	0.751	0.745	0.746	0.755
F1-Score	0.753	0.737	0.742	0.738
Precision	0.817	0.739	0.826	0.767
Recall	0.700	0.736	0.673	0.712
ROC AUC	0.828	0.824	0.824	0.852

Evaluating and comparing models in terms of common classification metrics using held-out data is an important way to determine their predictive capacities. It is, however, equally important to assess a model’s practical usefulness as well. Thus, the predicted crop area is compared to the actual crop area. **Table 3** shows the actual crop area as of the KAZA Land Cover 2020 dataset in comparison to the predicted crop area for the six Bengo regions according to random forest. There are substantial differences between predicted and actual crop area per region. These differences might stem from different sources, e. g. not enough training labels, not good enough quality of the training labels, oversampled crop class, variations in cropland distribution across regions or a too simplistic model. It is somewhat surprising to see such deviations considering the acceptable predictive performance according to the classification metrics shown above.

Table 3: Total area (KAZA Land Cover 2020) and predicted crop area for the six Bengo regions according to random forest.

	Actual crop area	Predicted crop area
Binga	3,498.84 km² (26.07%)	4,975.36 km² (37.07%)
Hwange	980.77 km² (3.58%)	9,498.24 km² (34.67%)
Mufunta	377.28 km² (5.86%)	471.37 km² (7.32%)
Mulobesi	36.67 km² (1.02%)	548.19 km² (15.25%)
Sichifulo	519.85 km² (14.61%)	942.04 km² (26.48%)
Zambezi	652.03 km² (3.81%)	3,555.69 km² (20.78%)

Unfortunately, predictions from the OpenMapFlow model are currently lacking for the six Bengo regions due to resource restrictions. Instead, predictions were made for a selection of smaller regions of interest, part of the ARISE sites, namely Sikunga in Namibia, Nyawa in Zambia and Kachechete, Chikandakubi, Chidobe and Nemananga in Zimbabwe. **Table 4** shows the actual crop area as of the KAZA Land Cover 2020 dataset as well as the predicted crop area according to random forest and the OpenMapFlow model. Again, the predictions deviate quite substantially from the actual values. In addition, there are inconsistencies among the models depending on the region of interest.

Table 4: Actual crop area (KAZA Land Cover 2020) and predicted crop area according to random forest and the OpenMapFlow model for three regions of interest located in Namibia (NAM), Zambia (ZAM) and Zimbabwe (ZIM).

	Total area	KAZA Land Cover 2020	Random forest	OpenMapFlow
NAM	287.59 km²	0.42 km² (0.15%)	25.00 km² (8.69%)	52.13 km² (18.13%)
ZAM	421.65 km²	154.28 km² (36.71%)	225.88 km² (53.57%)	231.16 km² (54.82%)
ZIM	722.82 km²	177.64 km² (24.66%)	243.03 km² (33.62%)	132.39 km² (18.32%)

Figures 3, **4** and **5** show the predicted crop maps for 2020 according to random forest (left) and the OpenMapFlow model (right). It is evident that the models agree in some areas while they disagree in others although both perform equally well in terms of all considered classification metrics across both the validation and the test set.

Figure 3: ARISE site NAM (Sikunga) - predicted 2020 crop map according to random forest (left) and the OpenMapFlow model (right).

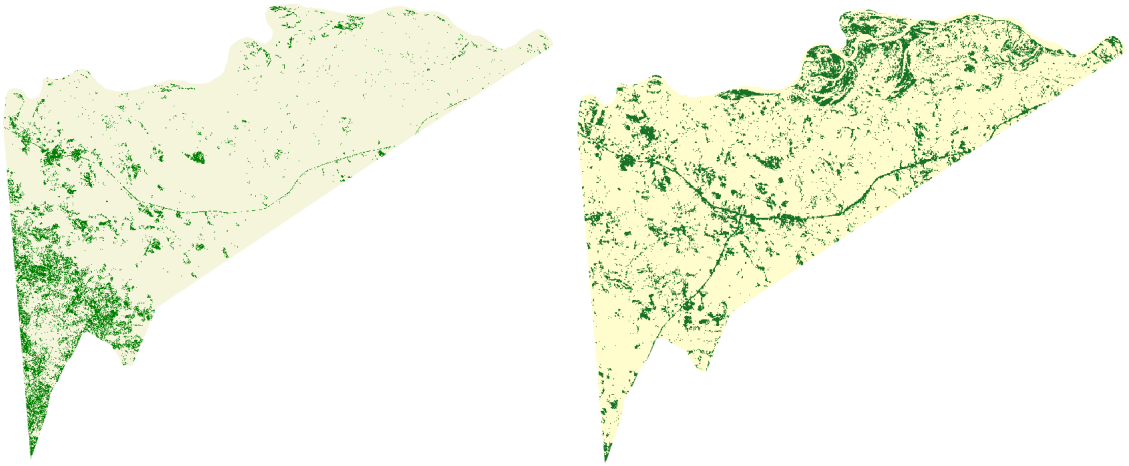


Figure 4: ARISE site ZAM (Nyawa) - predicted 2020 crop map according to random forest (left) and the OpenMapFlow model (right).

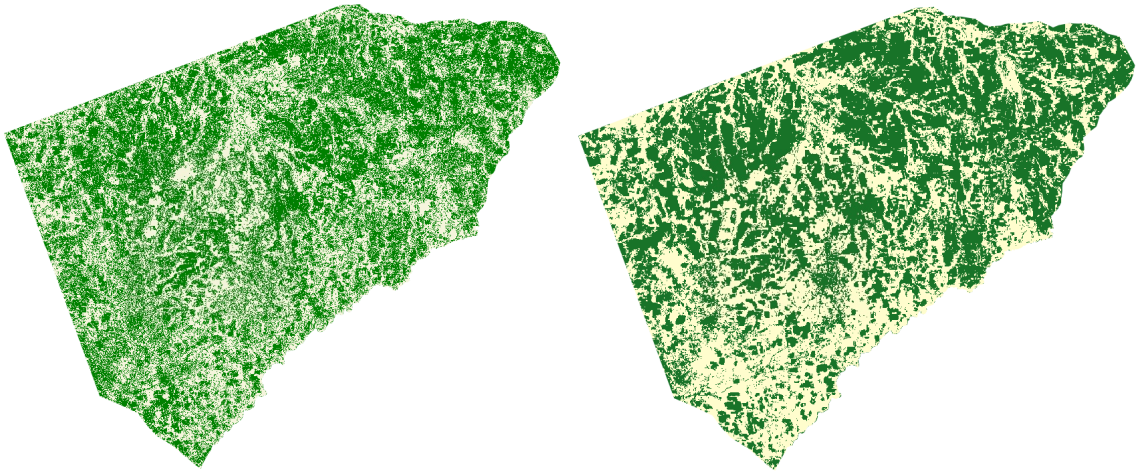
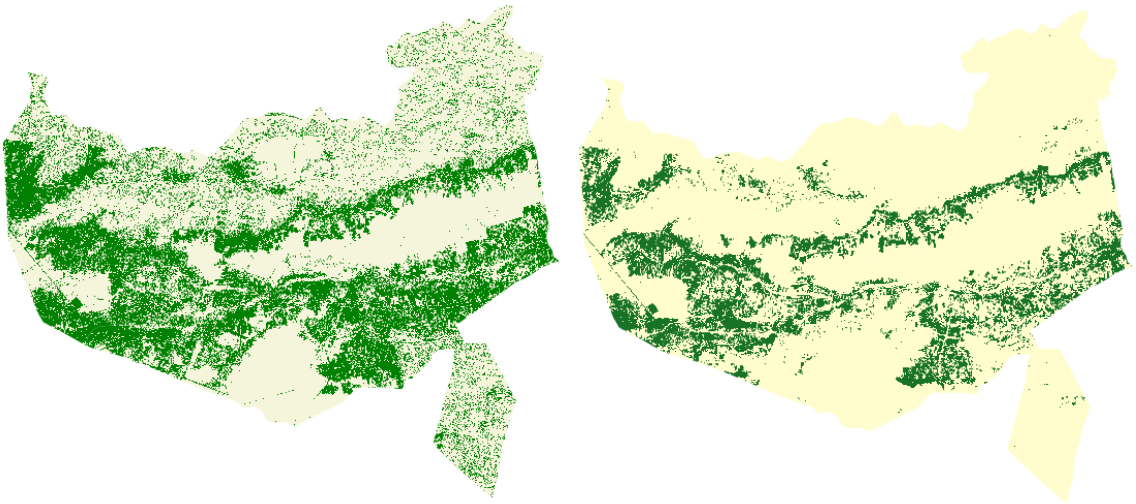


Figure 5: ARISE site ZIM (Kachechete, Chikandakubi, Chidobe and Nemananga) - predicted 2020 crop map according to random forest (left) and the OpenMapFlow model (right).



Conclusion and next steps

ML and publicly available satellite data can be used to create annual crop maps efficiently at scale to support impact monitoring efforts in KAZA, fostering sustainable agricultural practices that benefit people and wildlife. Although the first results seem promising, there remains room for improvement. The following tasks should be considered to improve the quality of the produced crop maps in order to best reach the desired impact:

- Cropland comprises only a small fraction of all land cover classes in KAZA, with regional clusters. The training labels were sampled in a way to produce a balanced dataset. It remains an open question what class ratio is ideal in the present case.
- The created crop maps need to be further evaluated and checked against high-quality ground truth data (e. g. recent data collected during a field campaign and geo-referenced using higher-resolution Planet data).
- It would be desirable to use higher-quality (and more) training labels to improve crop map development (e. g. recent data collected during a field campaign and geo-referenced using higher-resolution Planet data).
- Better model training and evaluation would enable a better decision-making process regarding model selection. Leveraging OpenMapFlow's capabilities requires financial resources for using Google Cloud.
- Finally, annual (e. g. 2020, 2021 and 2022) crop maps for the six Bengo regions could be created using the selected model and change detection with regard to cropland distribution could be performed.

References

KAZA: https://space-science.wwf.de/KAZAStory/

KAZA Land Cover 2020 dataset: https://space-science.wwf.de/KAZAlandcover/
NASA Harvest's OpenMapFlow: https://github.com/nasaharvest/openmapflow
Project GitHub repository: https://github.com/alexvmt/farm_plot_detection