Simple Random Forest and Satellite Images Time Series for Cropland mapping

ANTOINE SAGET

1 INTRODUCTION

In this brief technical report we present our simple solution (ranked 2nd place) to the GEO-AI Challenge for Crop Mapping by ITU that achieve a 0.943 accuracy on the private leaderboard. Our solution is based on a simple Random Forest classifier trained on Sentinel-2 time series data. Our contribution lies mostly in retrieving the Sentinel-2 time series data from Google Earth Engine. The preprocessing is kept to a minimum as the Random Forest classifier is robust to unclean data. All code to reproduce results is available at https://github.com/ITU-GeoAI-Challenge/2nd-place-GEO-AI-Challenge-for-Cropland-Mapping.

In Section 2 we explained how data is gathered and preprocessed, then in Section 3 and 4 we provide two simple studies respectively on the choice of time series length and period and on the choice of Sentinel-2 bands. Then, we briefly comment on our model choice in Section 5. Finally, we present and discuss our results in Section 6.

2 GATHERING SENTINEL-2 TIME SERIES FROM GEE

For each sample in the training and test sets, we gather Sentinel-2 L2A time series without cloud filtering starting from one year before the challenge period up to one year after the challenge period. Please note that careful consideration must be taken in the making of the Google Earth Engine request. Without care, it's easy to request only aligned time series, this will result in the loss of all unaligned timesteps and therefore a significant loss of information. Please refer to the implementation for more details, specifically part 1. of the full_study.ipynb notebook. We gather all bands from Sentinel-2 plus an additional NDVI band ($\frac{B8-B4}{B8+B4}$). This result in 3000 time series of different length and different starting and ending dates. The number of timesteps varies a lot between timeseries: from 100 up to 700+ timesteps. This variation occurs even within the same country. This is due to the fact that Sentinel-2 revisit time is not always 5 days due to tile overlapping and other factors.

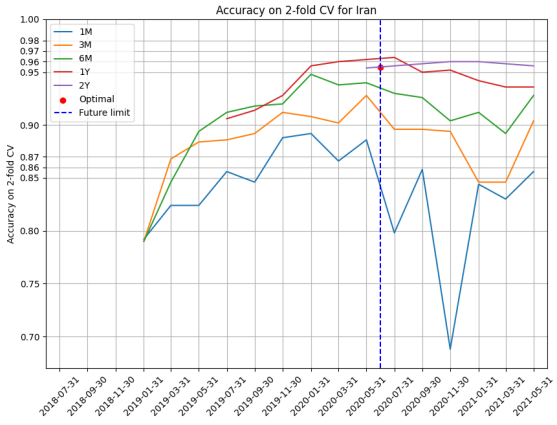
Random Forest does not support the data in it's current state. We first interpolate every missing timestep to fill every single day of each time series. Then we resample the time series to only keep every 5 timesteps as Sentinel-2 revisit time is usually 5 days. This result in per country aligned time series of equal length. When feeding the 2D time series of shape ($n\_timesteps$, n_bands) to the models, we first flatten them into a 1D vector of length $n\_bands * n\_timesteps$.

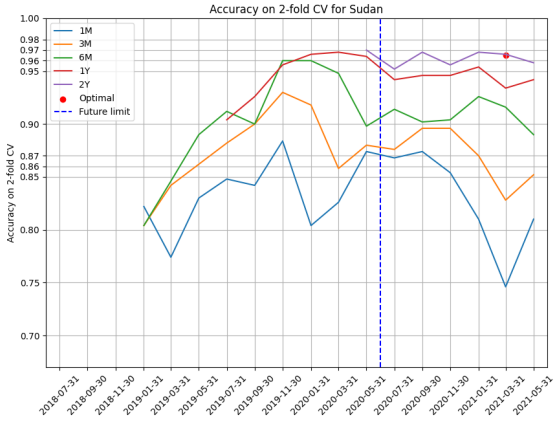
Considering the robustness of Random Forest we don't apply any further preprocessing such as cloud filtering, standardization or normalization. However this might improve the results, if done well, especially for models not as robust to outliers or for models sensitive to scale.

Author's address: Antoine Saget, antoinesaget19@gmail.com.

3 BRIEF STUDY ON THE IMPACT OF DIFFERENT TIME RANGES

In this section we study the impact of various time series lengths and periods (end dates) on cross-validation classification accuracy on the training set using a simple Random Forest. In practice we consider time series length of 4, 12, 24, 48, and 96 weeks. For each length we train Random Forest with cross validation on a number of periods (every two months). This gives us as an idea of the best time series lengths and periods for classification. Each country is trained separately. In Figure 1, we display the results. The vertical blue dashed lines indicate when we start using data beyond the given challenge time ranges (i.e. in the future).





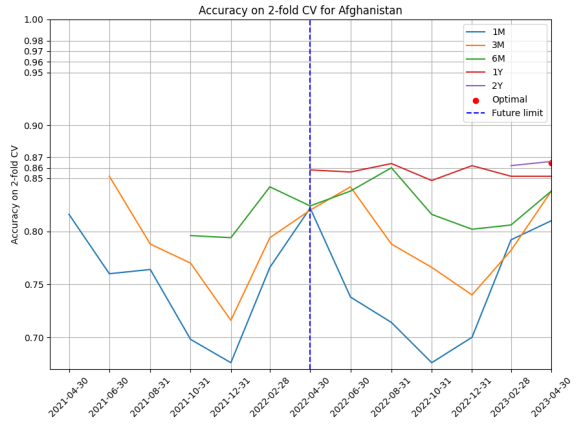


Fig. 1. Training set CV Random Forest Accuracy in function of Country (plot), timeseries lengths (color) and end-date (x axis).

From this experiment, we can see that the best results are obtained with longer time ranges and saturates at around 1 year (red line). We can also see that better result are obtained when using data around the given challenge time ranges (better accuracy is obtained within the few months around the blue line).

We can also see that Afghanistan has significantly worse results than the other countries with a max cross-validation accuracy of 0.87 compared to 0.97 for the other countries.

We choose optimal time ranges and time periods for each country, they are represented by the red dots on the figures. Choosing optimal parameters based only on the best accuracy might lead to overfitting. Thus, we choose them visually based on accuracy and stability in accuracy (how flat the line look) around the optimal time period. The chosen optimal time ranges and time periods are :

Country	Start Date	End Date	Length
Sudan	2019-05-29	2021-03-31	2 years
Afghanistan	2021-06-27	2023-04-30	2 years
Iran	2018-08-28	2020-06-30	2 years

4 BRIEF STUDY ON THE IMPACT OF SENTINEL-2 BANDS CHOICE

In this section we investigate which Sentinel-2 bands are the most useful. We train a Random Forest with the same parameters on different band combinations and compare the results. Considering that Random Forest can ignore irrelevant features, our model will not suffer much if we give it too much features. However using all bands is not necessarily the best option as this will increase training time and memory usage.

We can see in Table 1 that even with only the NDVI band, we can achieve a high cross-validation accuracy of 0.93. All but the "SCL only" configurations perform well so the user can easily limit the number of bands to significantly reduce training time and memory consumption without much loss in classification accuracy (see Appendix A.1). For the challenge submission, we choose [B2, B3, B4, B8, Lon, Lat, NDVI, SCL] (SCL is a categorical feature, so it's one-hot encoded). We've decided to keep Lon, Lat and SCL despite no significant accuracy improvement as they provide information different in nature than radiometric bands and might be useful to the model.

5 MODEL

In this study, we didn't explore different models. Considering our very simple preprocessing, our hypothesis is that most models will be outperformed by Random Forest (or Random Forest based models) and that it will require careful data cleaning for other methods to surpass Random Forests.

For the challenge submission, we took the default python scikit-learn Random Forest with all default parameters except the max depth that we set (arbitrarily) at 10 (shallow forest) to prevent the model from overfitting. One separate Random Forest is trained per country.

6 DISCUSSION AND RESULTS

In this report we have shown that a simple Random Forest model paired with mostly unprocessed Sentinel-2 time series can achieve a high accuracy on the GEO-AI Challenge for Cropland Mapping by ITU challenge.

Using the optimal time series lengths and periods found in Section 3 with the bands selected in Section 4 trained on one model per country as discussed in Section 5, we achieve 0.943 accuracy on the private leaderboard.

The strengths of our solution are:

- Simple model that is easy to train
- Robust to unclean data (data is almost used raw from GEE)
- Can easily scale to larger datasets
- Can be trained easily on CPU in a few minutes on Google Colab (< 3min to reproduce final results and <15min to reproduce the full study)
- Can be trained on GPU if dataset is very large (with NVIDIA Rapids cuML Random Forest implementation for example)
- Can be adjusted to use less memory by reducing the number of timesteps or bands without significant accuracy loss (see Appendix A.1)

The weaknesses of our solution are:

- The simple preprocessing might not be enough for other models, especially models sensitive to outliers or scale. For example standardization or normalization might be necessary for Neural Networks.
- The GEE download is long. Optimizing the GEE pipeline might be necessary for larger datasets if time is an issue.
- The data interpolation and reindexing is long when compared to the training time of the model. Further optimization might be necessary for larger datasets if time is an issue.
- Currently the best results are obtained using future data. If the model is meant to be used in real-time, only past data can be used. This can lead to a very slight decrease in accuracy from 0.94 to 0.93 (see Appendix A.2).
- Currently, there is one model per country, meaning that for example, Afghanistan and Sudan data is never used during

B2	B3	B4	B5	B6	B7	B8	B8A	B11	B12	SCL	NDVI	Lon	Lat	Accuracy
										V				0.901
											V			0.933
	V					V								0.935
	V	V				V								0.937
V	V	V				V								0.932
V	V	V				V				V	V			0.939
V	V	V				V				V	V	V	V	0.933
V	V	V	V	V	V	V	V	V	V	V	V			0.935

Table 1. Table of tested Sentinel-2 L2A bands configurations and resulting CV Accuracy on training test.

the training of the Iran model. We show in Appendix A.3 that accuracy is slightly worse with one model for all country instead of one model per country. However, this is a waste of data that could be exploited by finding ways to align the signal (in both time and radiometric response dimensions) between the countries for example.

- Currently the model is weaker on Afghanistan than on other countries. Further investigation is necessary to understand why.

Possible improvements:

- Ensembling multiple RFs with different parameters (such as a wide, a shallow and a normal RF) might improve the results.
- Adding Land cover classification from the Dynamic World V1 collection as an additional feature might improve the results. This collection is already the result of a classification but from a model trained with 5 billion pixels of training data.
- As shown in [Rußwurm et al. 2019], transformer models outperform RFs on classification with Sentinel-2 time series. This might be a good improvement to try over our current model provided that preprocessing is done correctly. However this will require significantly more work to implement compared to a simple Random Forest.
- Finally, considering the amount of Sentinel-2 unlabeled data available, a semi-supervised approach might be a good improvement to try. Self-supervised pretraining on a large quantity of unlabeled data followed by finetuning on the labeled data for example. However, this will be more difficult to implement with longer training time and more memory usage (probably requiring a GPU).

REFERENCES

Marc Rußwurm, Sébastien Lefèvre, and Marco Körner. 2019. BreizhCrops: A Satellite Time Series Dataset for Crop Type Identification. *CoRR* abs/1905.11893 (2019). arXiv:1905.11893 http://arxiv.org/abs/1905.11893

A EXTRA RESULTS

A.1 NDVI only vs. [B2, B3, B4, B8, Lon, Lat, NDVI, SCL]

Bands	Acc on private LB	Time AMD Ryzen 9 5900x	Time Google Colab	X_train size
B2, B3, B4, B8, LON, LAT, NDVI, SCL (challenge submission)	**0.943**	5.1s	18s	31.1 MB
NDVI only	0.938	**3.2s**	1**5s**	**1.8 MB**

Table 2. Accuracy, training+prediction time and memory footprints when using NDVI only vs. bands selected in the challenge submission.

With only NDVI band, the drop in classification accuracy is minor but the memory footprint of dataset is divided by more than 17.

A.2 Using future vs. only passed data

Data	Acc on private leaderboard
Future and past data (challenge submission)	**0.943**
Past data only	0.932

Table 3. Private leaderboard accuracy on past data only vs. when allowing the use of some data coming from the future.

Without the use of future data, there is a minor drop in classification accuracy.

A.3 One model per country vs. One model for all countries

Model	Acc on private leaderboard
One model per country (challenge submission)	**0.943**
One model for all countries	0.934

Table 4. Private leaderboard accuracy when using one single model vs. one model per country.

Using one single model instead of one model per country lead to a slight decrease in classification accuracy.