GEO-AI Challenge for Cropland Mapping

Antoine Saget

november 23th 2023

The Challenge

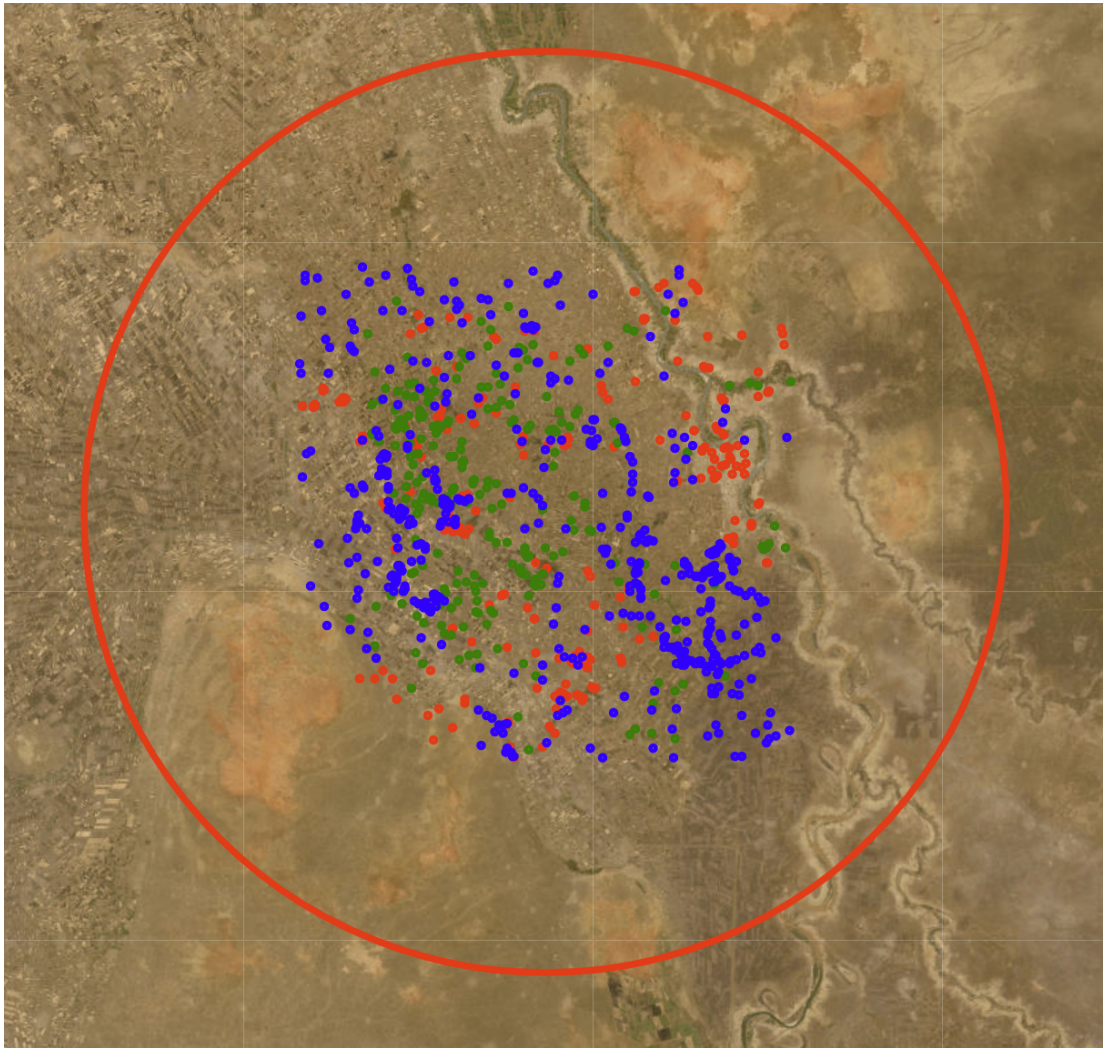


Figure 1 – Soudan dataset. Green = crop, Red = non-crop, Blue = test samples.

The Challenge

(a) Example of a crop area

(b) Example of a non-crop area

Solution overview

- Sentinel-2 timeseries
- One Simple Random Forest model per country

Data gathering - Sentinel-2 Timeseries

Sentinel-2 revisit time is 5 days or less and there might be missing timesteps (due to cloud for example)

Consequence : timeseries are not aligned and have different lengths

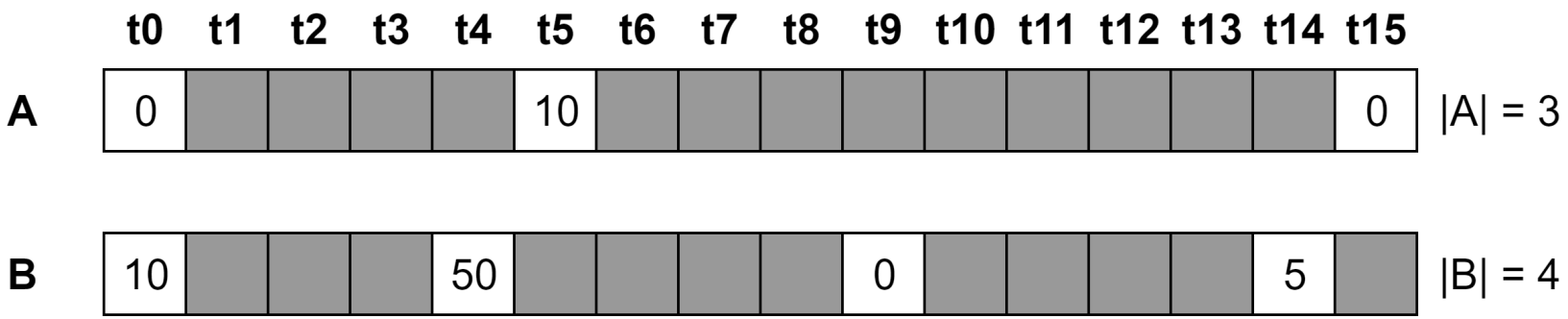


Figure 3 – Example of two unaligned timeseries of different lengths.

Data gathering - GEE Request

- No cloud filtering
- Add a Normalized Difference Vegetation Index (NDVI) band

Data gathering - GEE Request

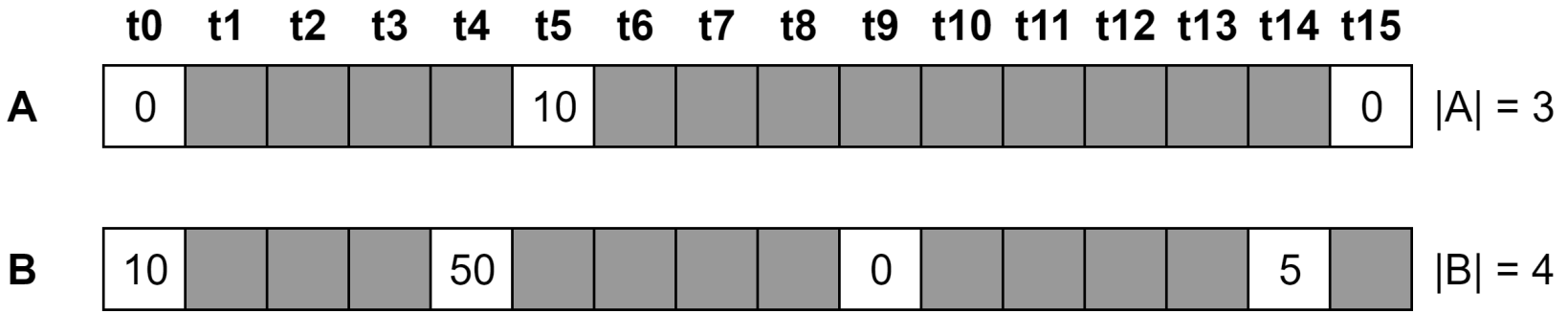


Figure 4 – Example of two unaligned timeseries of different lengths.

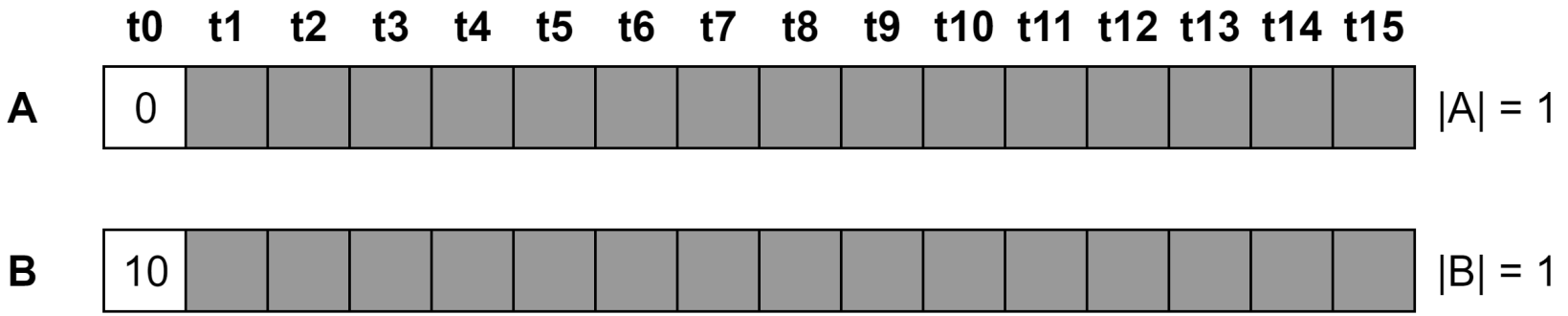


Figure 5 – Example of result given by a simple GEE request. Only matching timesteps are returned. Most of the data is lost.

Data preparation - 1. Interpolation

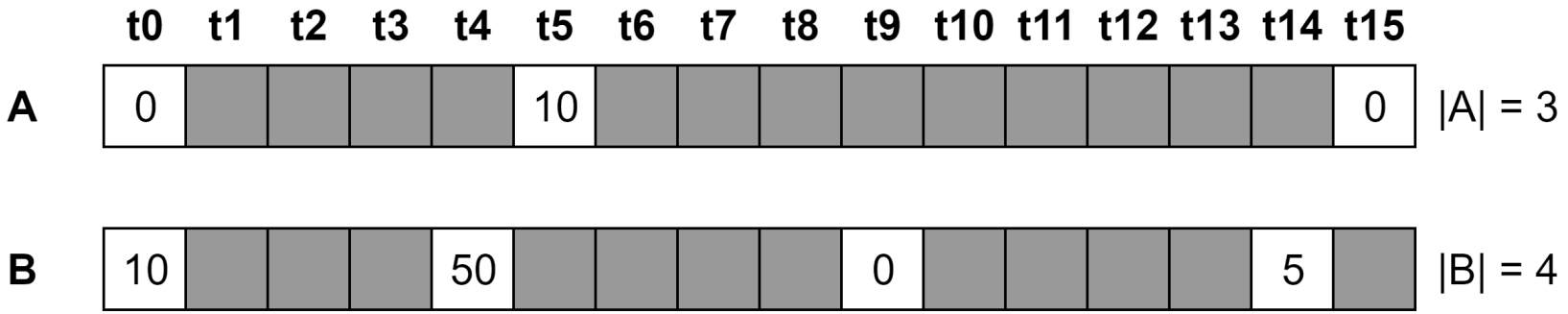


Figure 6 – Two unaligned timeseries of different lengths downloaded from GEE.

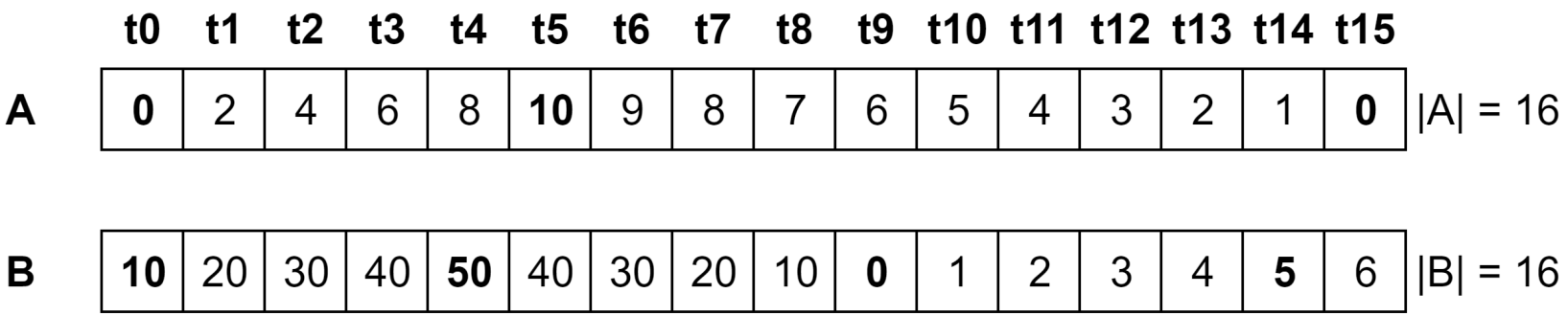


Figure 7 – After interpolation, the two timeseries are aligned and have the same length.

Data preparation - 2. Resampling

	t0	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14	t15	
A	**0**	2	4	6	8	**10**	9	8	7	6	5	4	3	2	1	**0**	|A| = 16
B	**10**	20	30	40	**50**	40	30	20	10	**0**	1	2	3	4	**5**	6	|B| = 16

Figure 8 – After interpolation, the two timeseries are aligned and have the same length.

	t0	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14	t15	
A	**0**					**10**					5					**0**	|A| = 4
B	**10**					40					1					6	|B| = 4

Figure 9 – After resampling, the two timeseries are shorter.

Model

Simple Random Forest implementation from scikit-learn :

- Fast (GPU implementation possible with cuML for large datasets)
- Easy to use
- Easy to implement
- Widely used in remote sensing
- Very robust, does not require much data preparation :
  - No need to normalize data
  - No need to remove correlated features
  - No need to remove outliers

Hyperparameter tuning

- Optimal time series length : 1-2 years
- Optimal time series dates : a few months before (past) + a few months after (future) the target date

Past + future data vs. past data only

Data	Accuracy
Future and past data (challenge submission)	**0.943**
Past data only (realtime scenario)	0.932

Table 1 – Private leaderboard accuracy on past data only vs. when allowing the use of some data coming from the future.

Best radiometric bands

Bands	Acc	Time	Mem X_train
2,3,4,8,LON,LAT,NDVI,SCL	**0.943**	5.1s	31.1 MB
NDVI only	0.938	**3.2s**	**1.8 MB**

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

Strengths of the solution

- Simple model, fast, easy to train ($<$ 3min to reproduce on Google Colab)
- Robust to unclean data
- Can easily scale to larger datasets
- Users can adjust to their need : accuracy, speed, memory consumption

Weaknesses of the solution

- The simple preprocessing might not be enough for other models
- The GEE download and the data interpolation are long when compared to the training time of the model. Further optimization might be necessary for larger datasets if time is an issue.
- Slight decrease in accuracy if only past data can be used (real time scenario).
- Currently the model is weaker on Afghanistan than on other countries. Further investigation is necessary to understand why.

Conclusion

High accuracy is achieved with a simple model and simple preprocessing.

I don't think much improvement can be made with the model itself nor the preprocessing.

I think the current bottleneck is the dataset and improvement can be made by expanding it (more labels) or cleaning it (correct labels if wrong).

One model per country vs. one big model for all countries

Model	Accuracy
One model per country (challenge submission)	**0.943**
One model for all countries	0.934

Table 3 – Private leaderboard accuracy when using one single model vs. one model per country.

Best timeranges and timeperiods

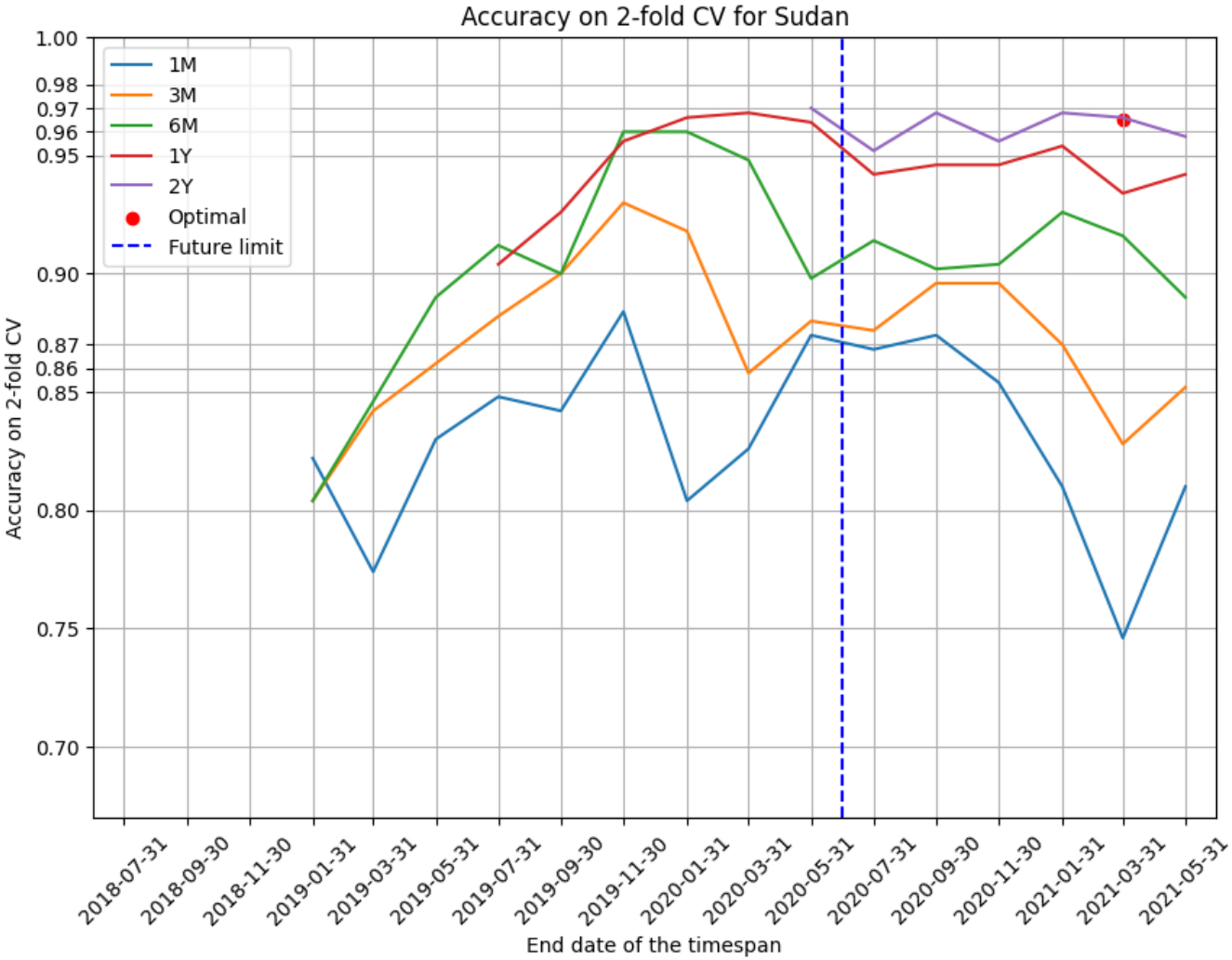
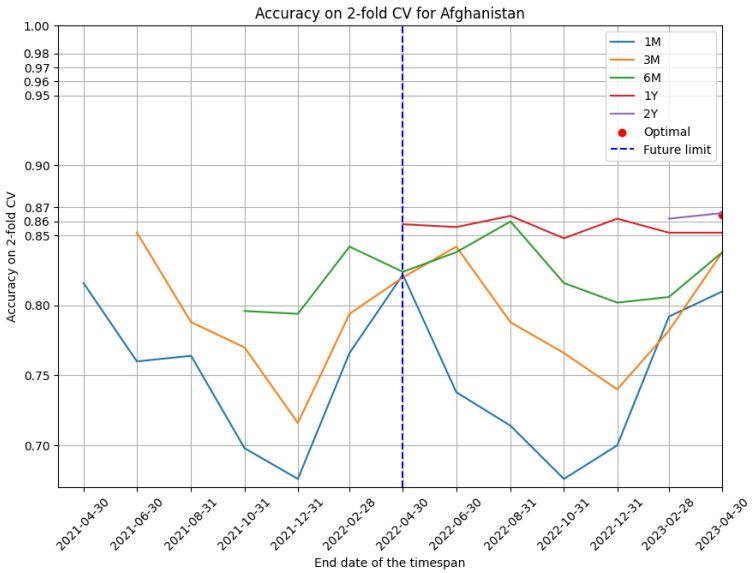
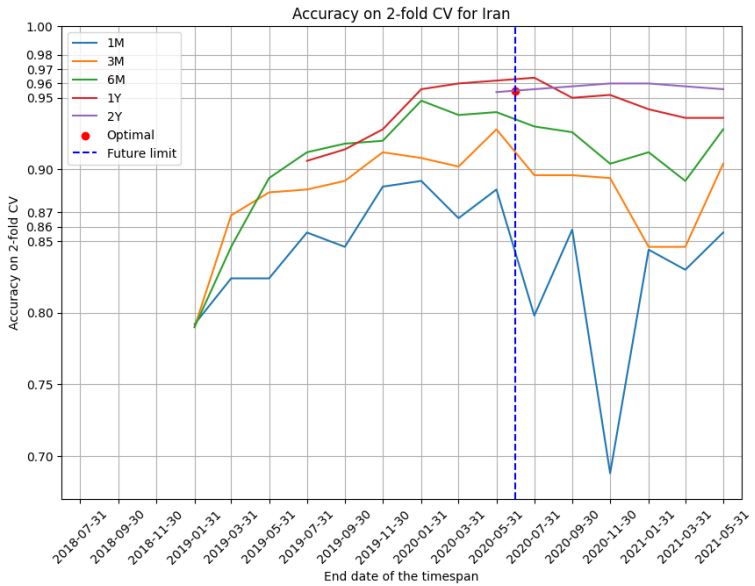


Figure 10 – Sudan RF CV Accuracy for different timespans and time periods.

Best timeranges and timeperiods



(a) Afghanistan RF CV Accuracy for different timespans and time periods.



(b) Iran RF CV Accuracy for different timespans and time periods.