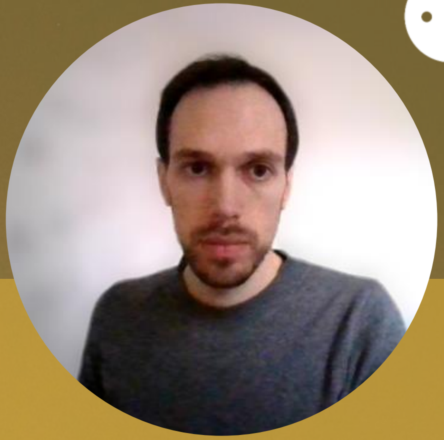
WorldCereal

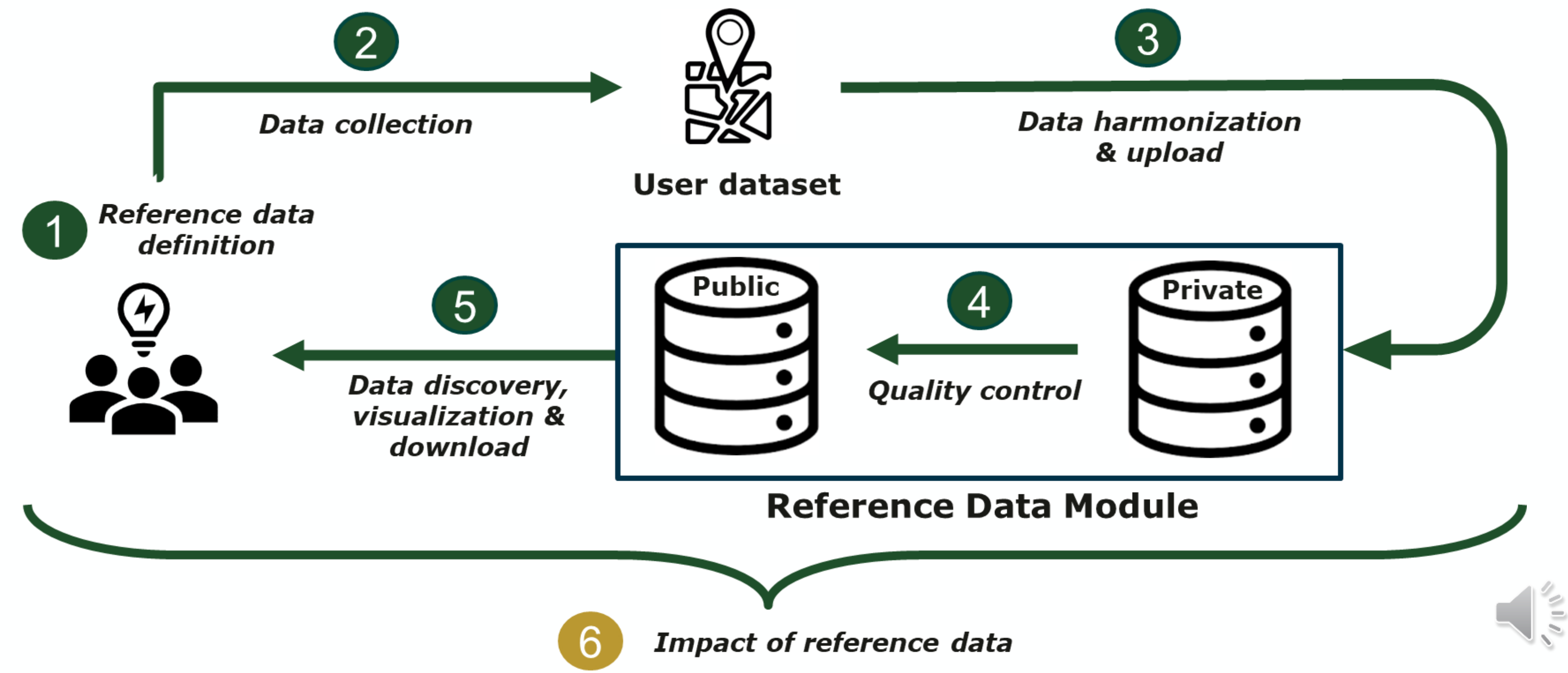
WorldCereal MOOC I



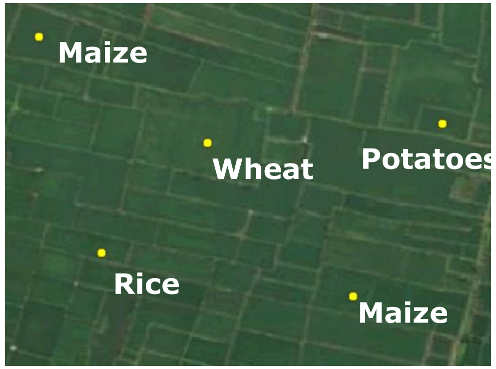
Impact of reference data on crop mapping

Jeroen Degerickx,
Christina Butsko (VITO)

MOOC I: Outline



Crop mapping from space?



Reference data

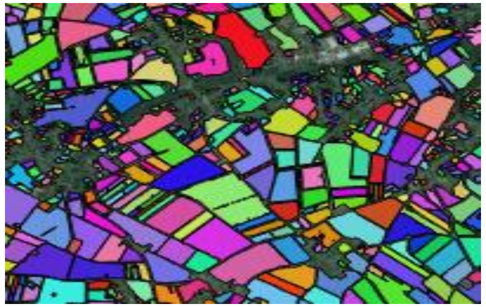
Sentinel-1 | Sentinel-2 | CopDEM | AgERA5

Time series over entire growing season
Satellite observations, meteorological data, altitude

PRESTO

Lightweight, Pre-trained Transformer for Remote Sensing Timeseries
Extracts general-purpose features useful for diverse EO applications

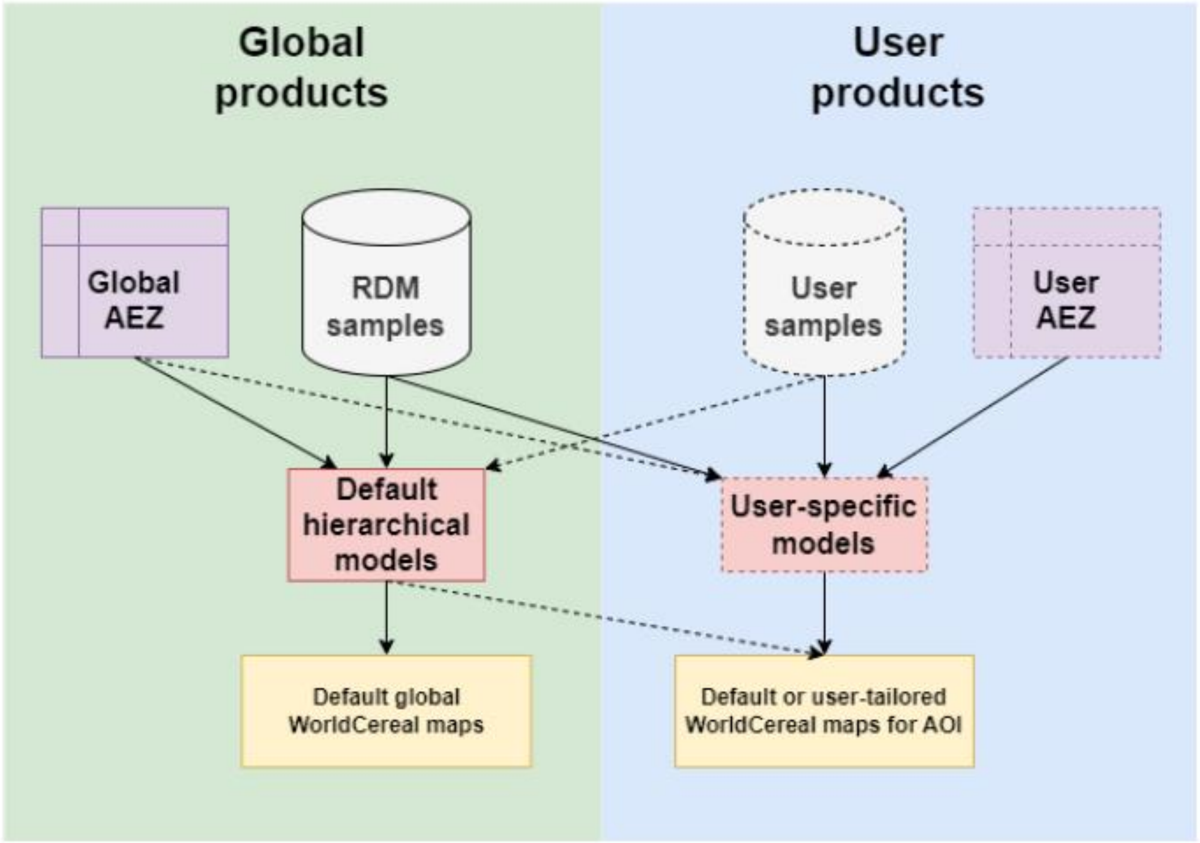
Crop identification model
Supervised pixel-based CatBoost classifier



Crop type map

WorldCereal

WorldCereal system: training your own models



Typical questions…

How much reference data do I need?

I already have data for another year, do I also need data for the year to be mapped?

I already have data for some crops, but want to add a new crop. How much more data do I need?

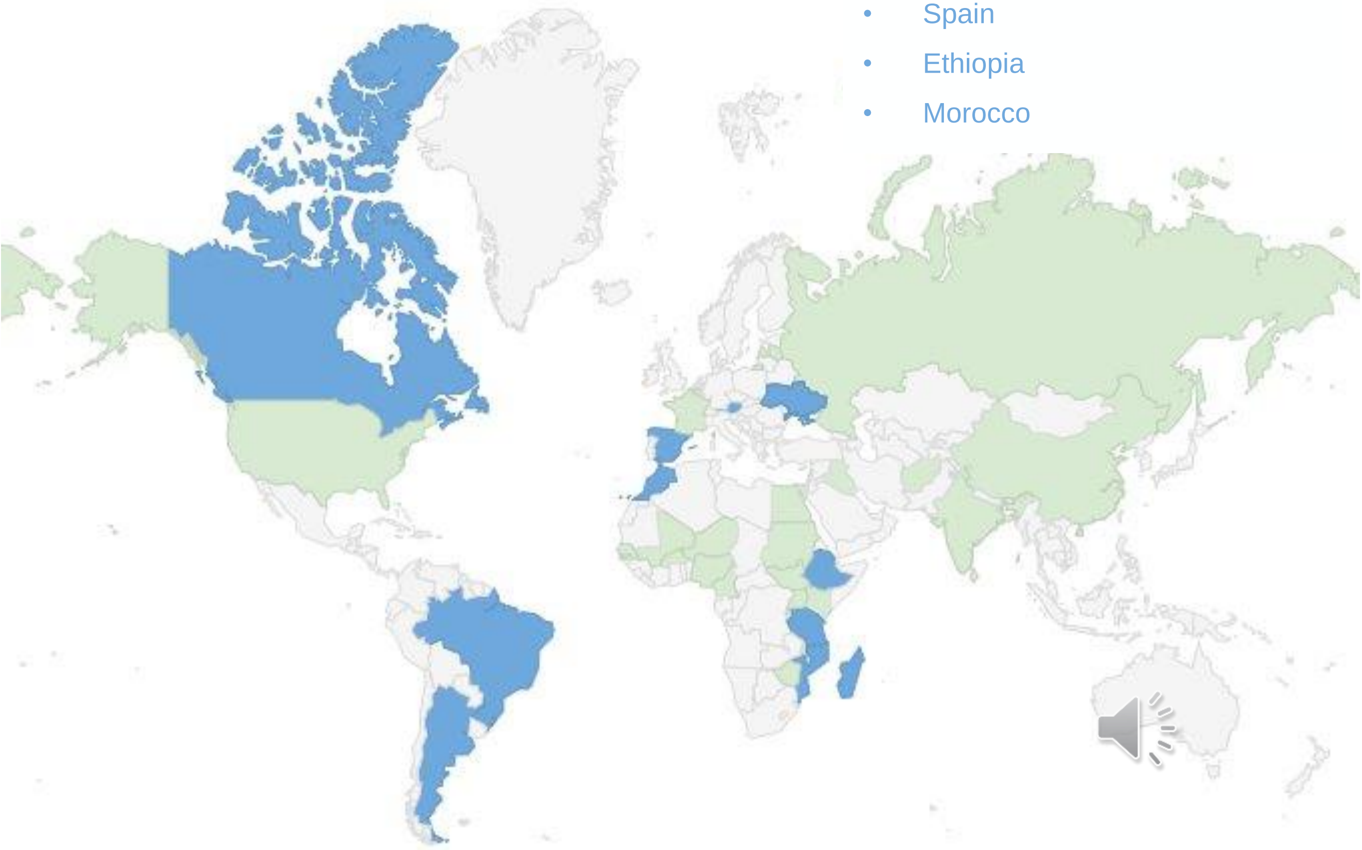
WorldCereal

Model performance versus data availability

Experimental setup

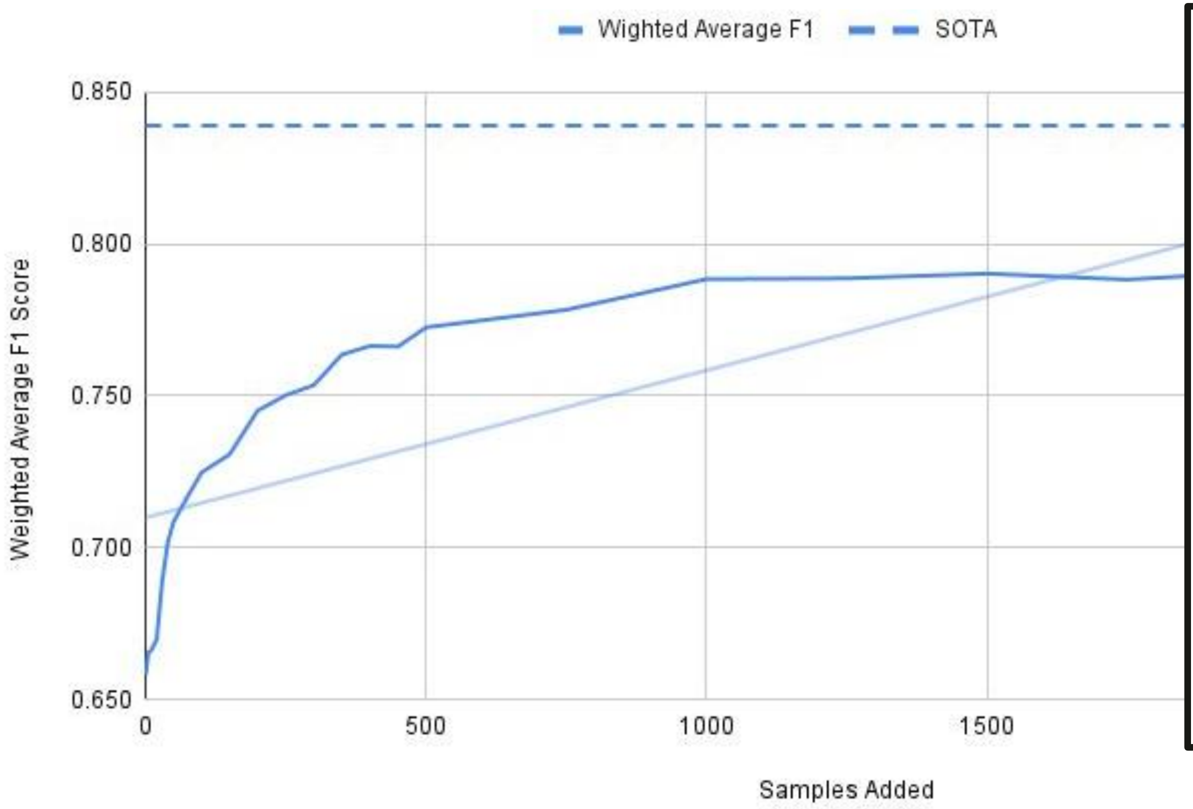
- Harvest all WorldCereal reference data currently available
- Train crop type models using all data, for 9 crops (wheat, barley, rye, maize, millet_sorghum, rapeseed, soybeans, sunflower, other) = state-of-the-art (SOTA)
- Set aside data from 11 countries (blue)
- Train a baseline model with all remaining data
- Gradually add training data from the left-out countries using various injection sizes (5, 10, 25, 50, 100, 250, 1000) and re-train model

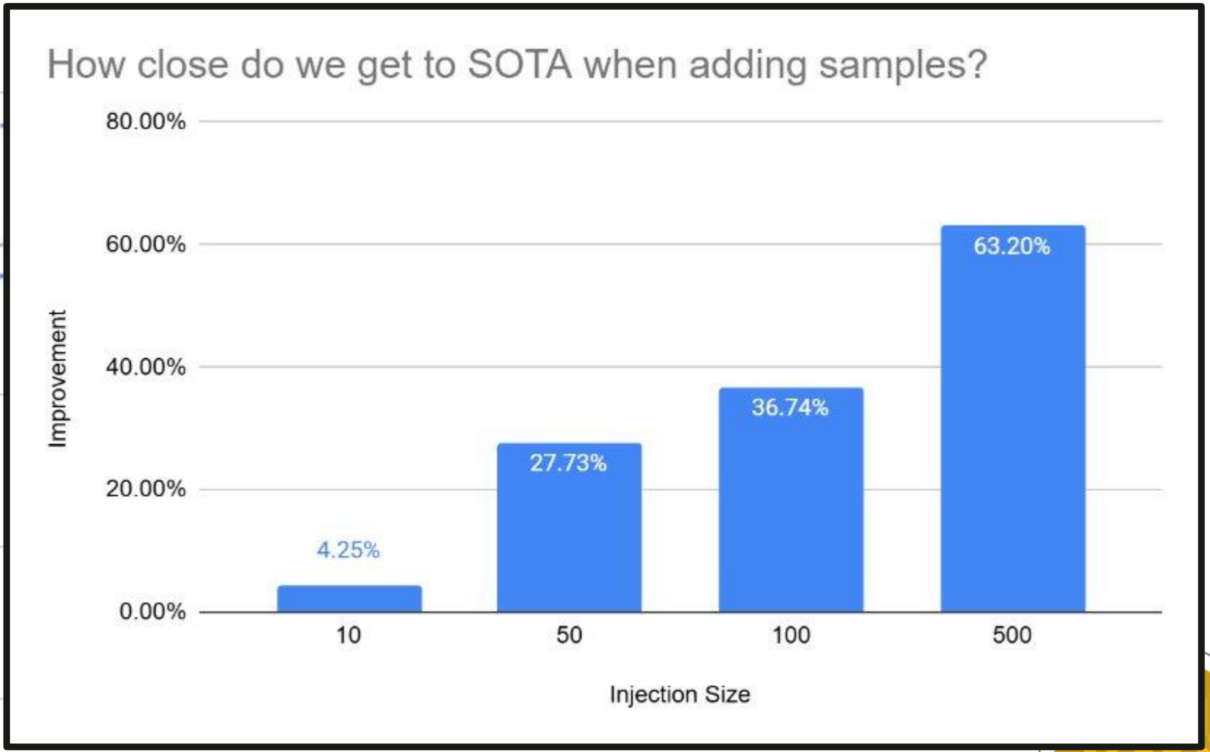
- Argentina
- Austria
- Brazil
- Canada
- Spain
- Ethiopia
- Morocco
- Madagascar
- Mozambique
- Tanzania
- Ukraine



Expected benefit of local reference data

Average increase in model performance when adding local data

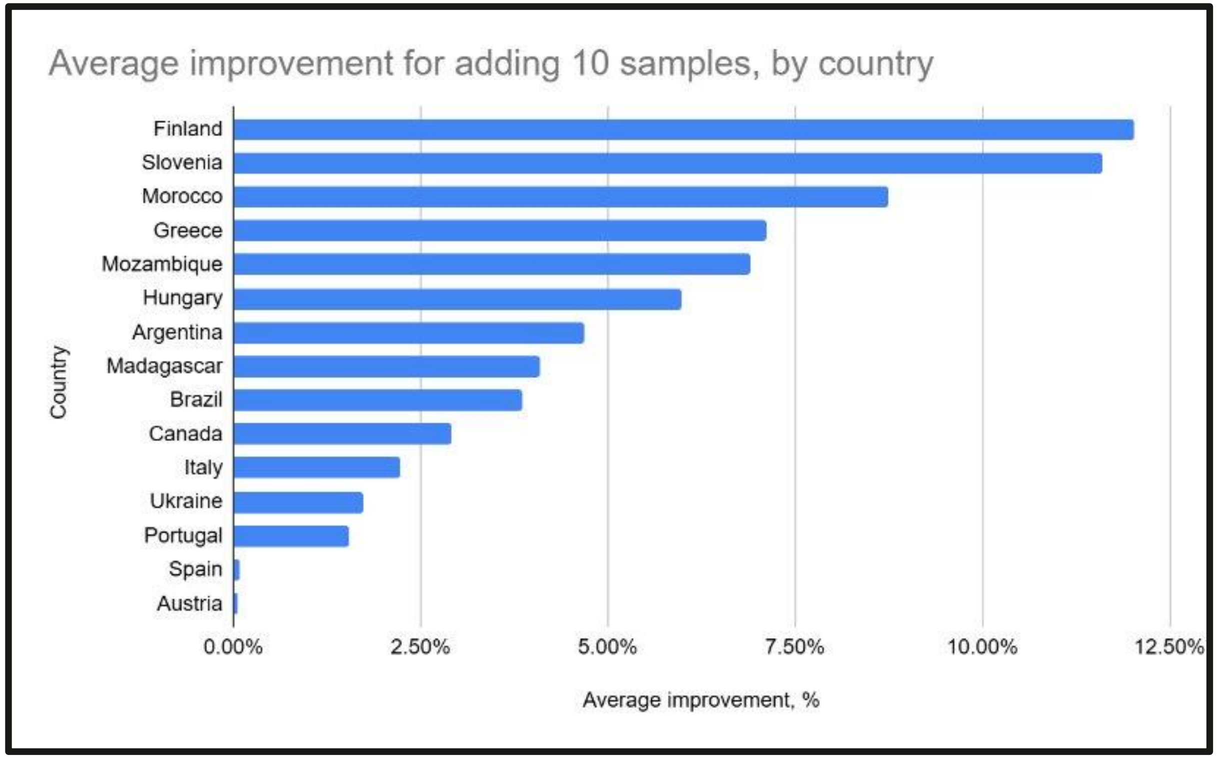


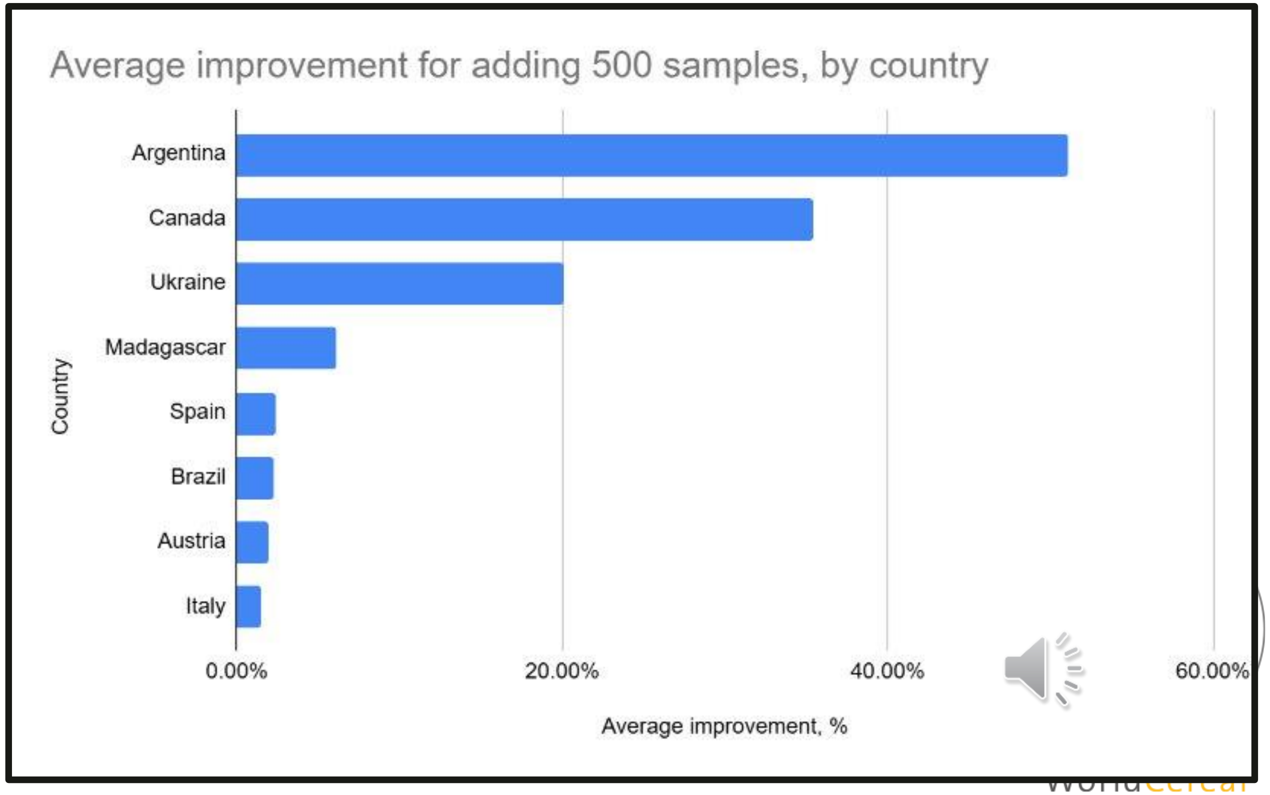


WorldCereal

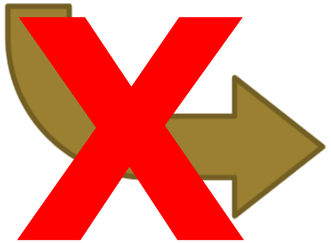
Expected benefit of local reference data

The magnitude of improvement varies significantly between countries and depends on the injection size





Note on model transferability

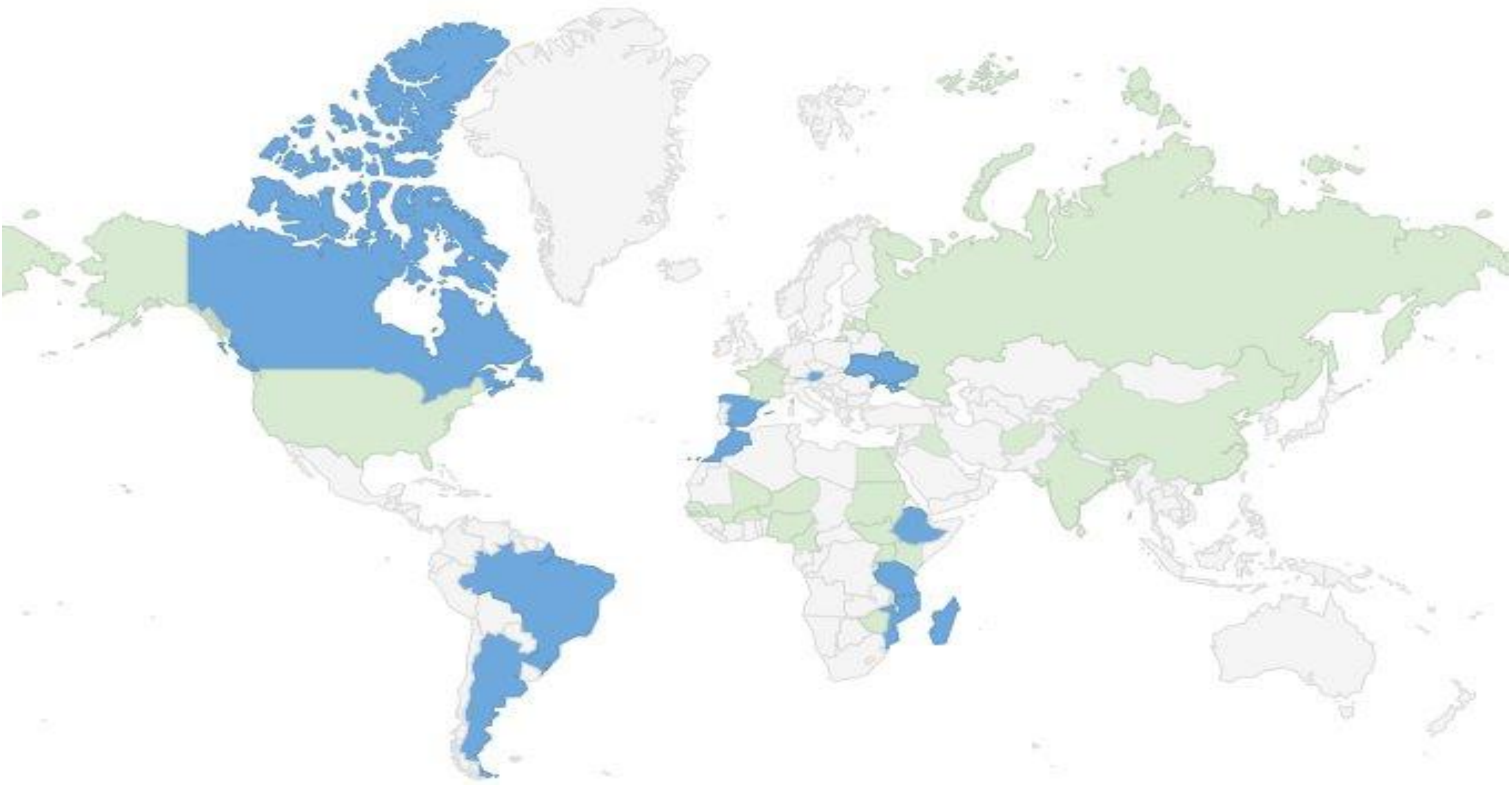


Factors to consider:

- Climate conditions
- Soil conditions
- Dominant crop types
- Agricultural management practices (field sizes, management activities like irrigation, timing of growing seasons)
- Other landscape features (altitude)

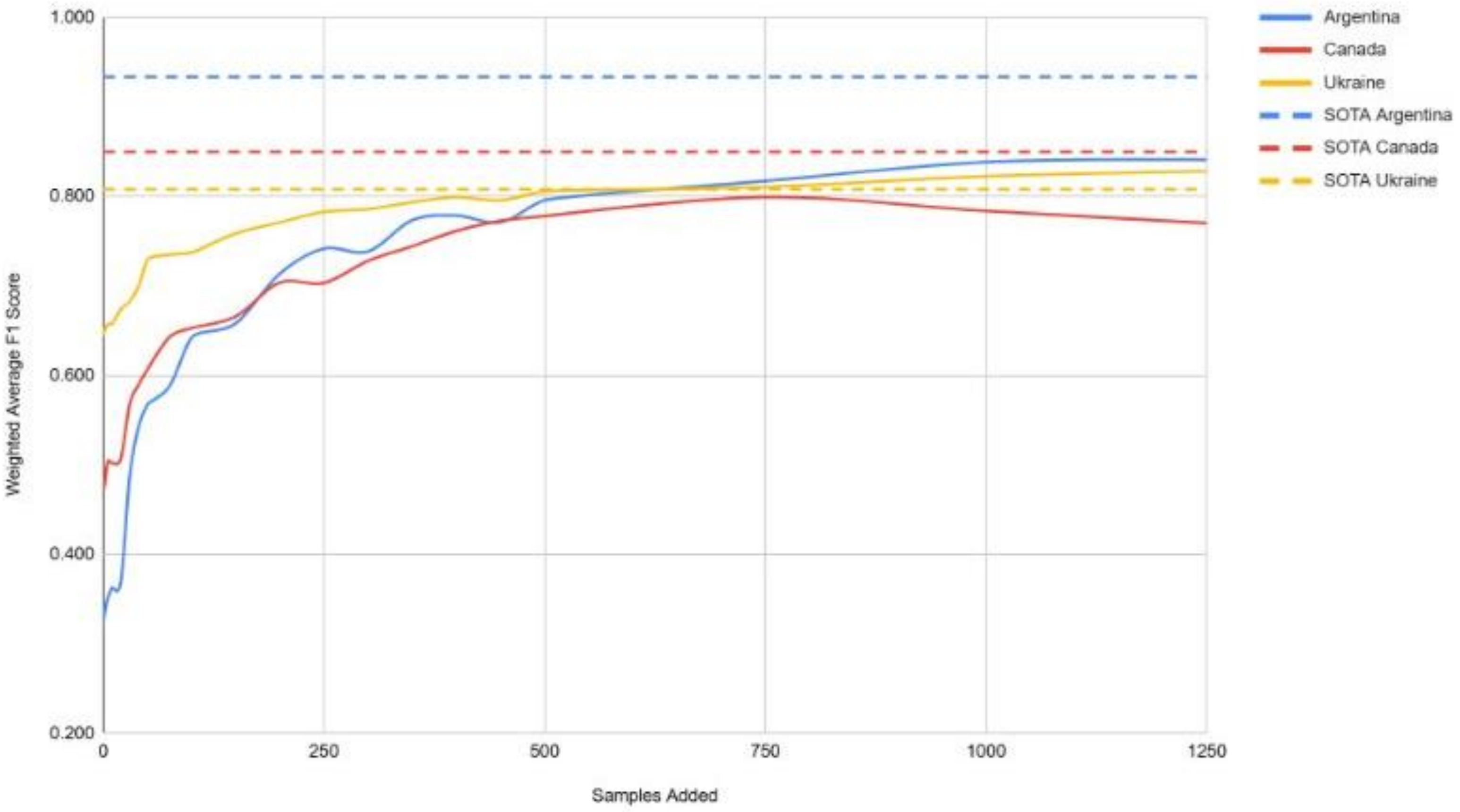
WorldCereal

Data-rich versus data-poor regions



WorldCereal

Data-rich versus data-poor regions

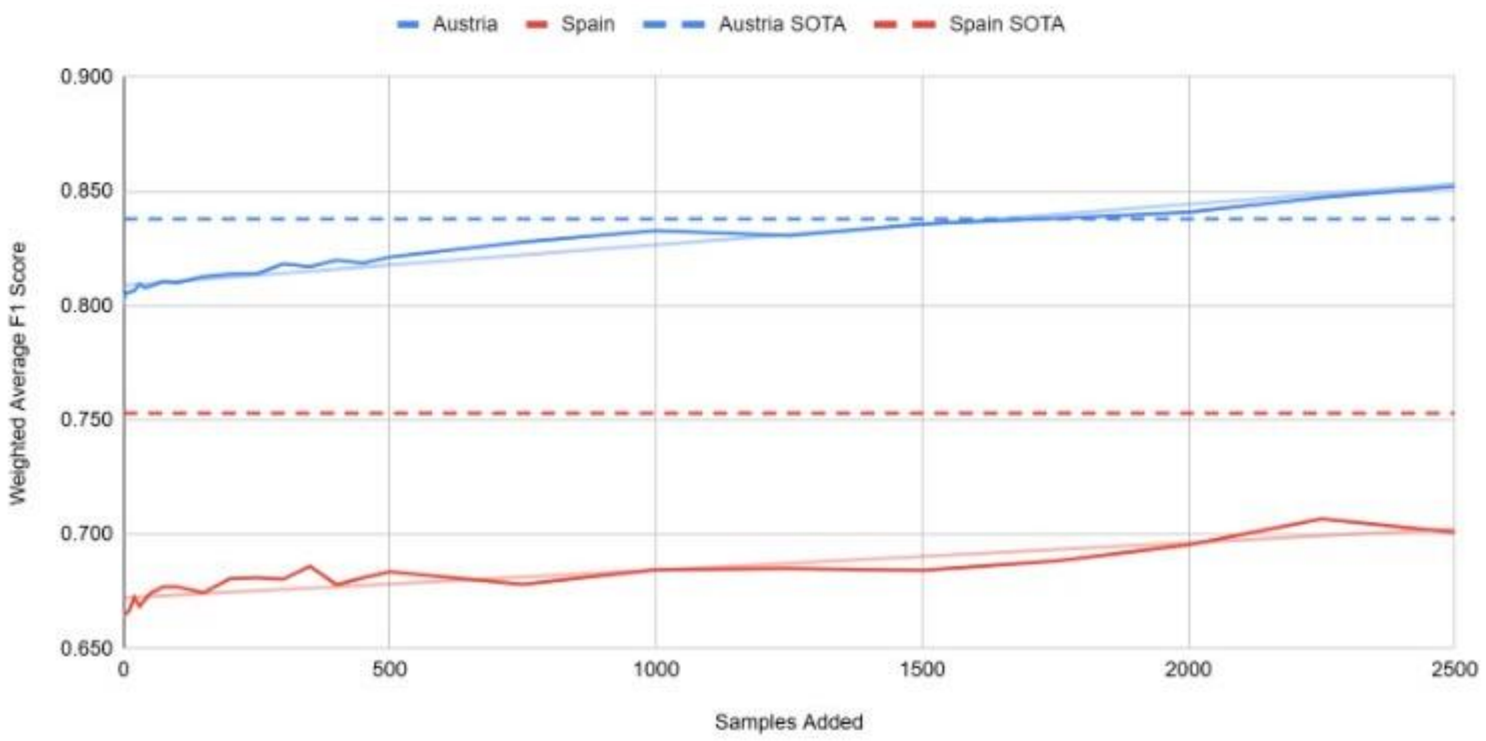


Three countries with large datasets demonstrate the **critical value of local data:**

- Argentina, Canada, and Ukraine all benefit significantly from local data injections, with substantial performance improvements.
- **Performance approaches SOTA with ~1,000 samples**; All three countries show diminishing returns beyond this point, indicating a saturation effect.
- **Regional differences in improvement:** Ukraine achieves close-to-SOTA performance with fewer samples (~250), while Argentina and Canada require larger injections to reach comparable levels.

WorldCereal

Data-rich versus data-poor regions



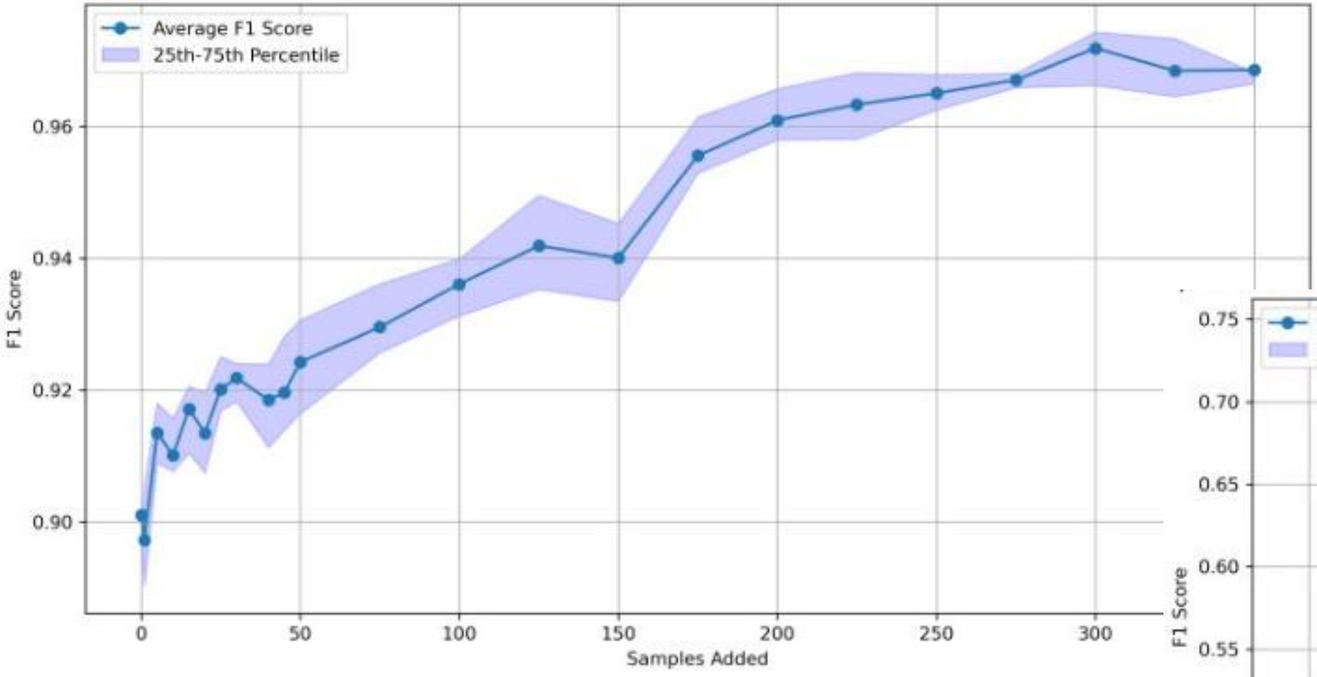
In data-rich regions, the incremental benefit of adding more local data diminishes:

- **High Baseline Performance:** In data-rich regions like Europe, models exhibit a high initial F1 score even without additional local data injections.
- **Smaller Returns from Additional Local Data:** Adding more local data in such regions results in marginal improvements in classification accuracy.

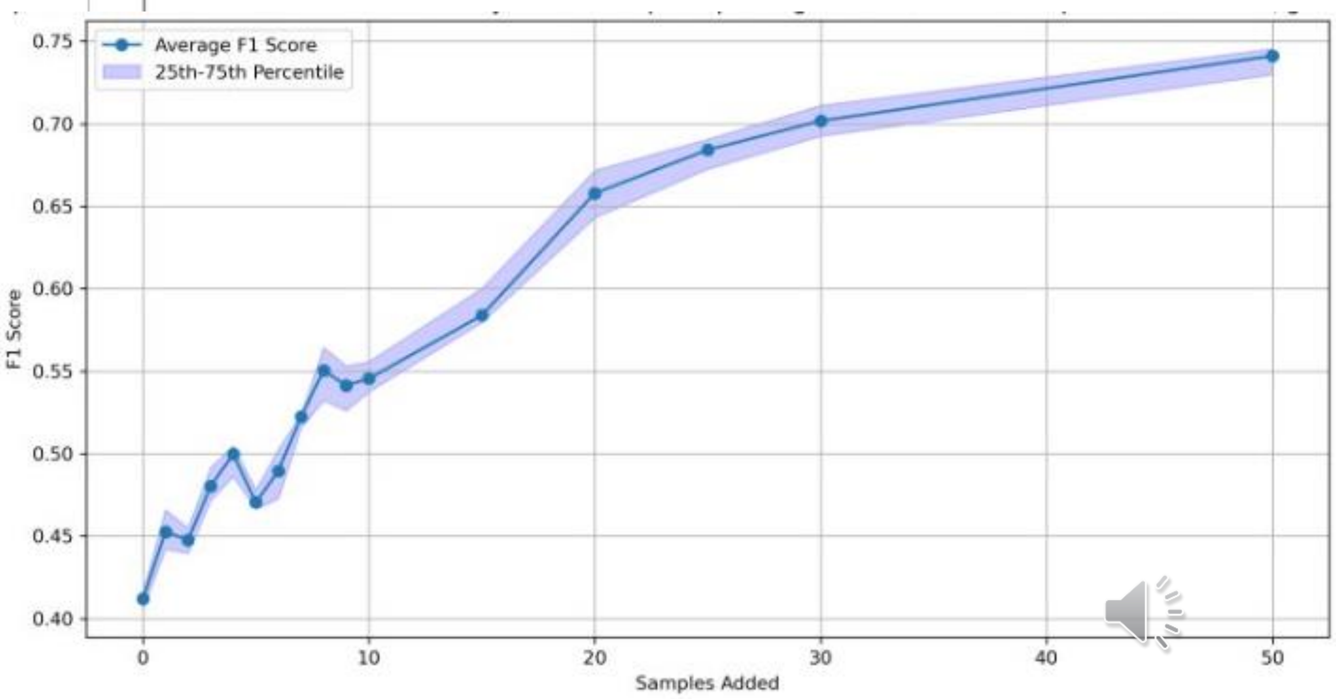
WorldCereal

Case 1: No data available for my region of interest

Brazil



Mozambique

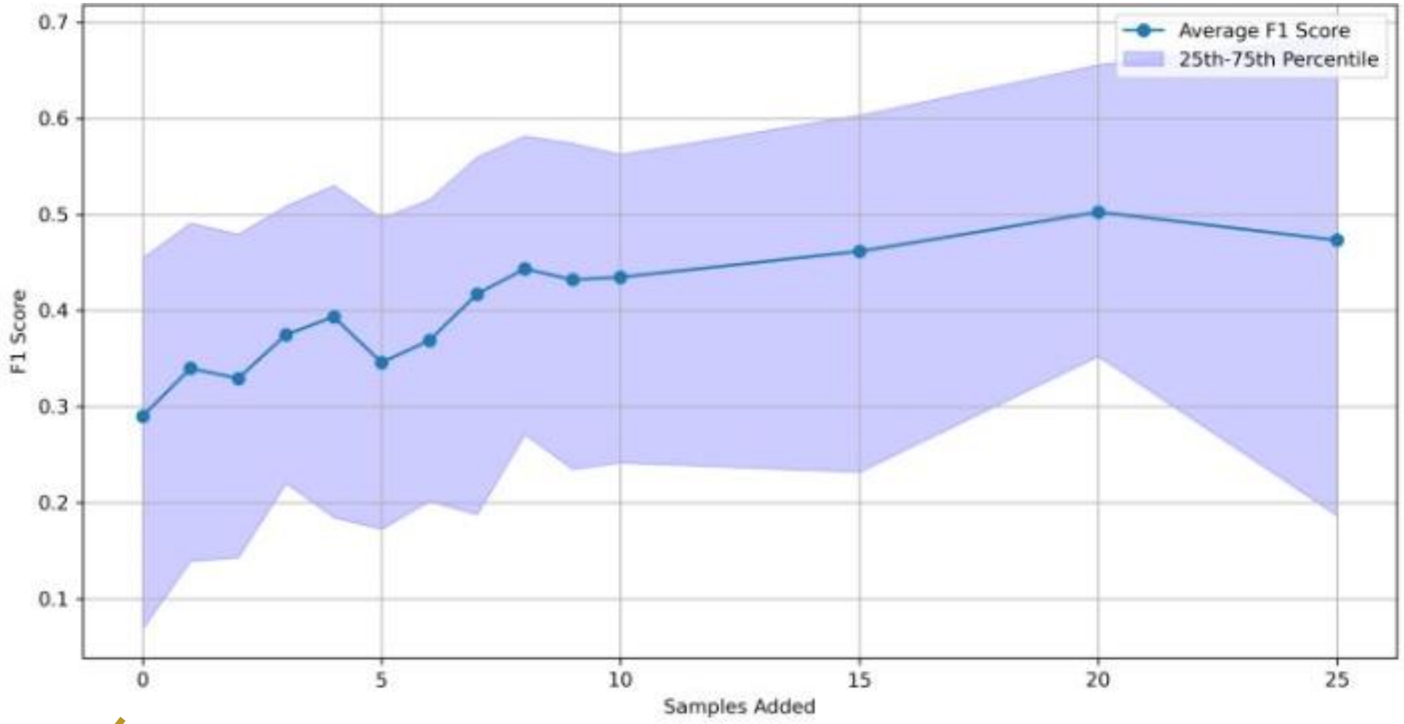


BR: Strong baseline; adding just 50-100 samples severely boosts performance!

MB: Lower baseline, more challenging region; substantial improvement can be achieved with only 50 samples! More data will be needed…

Case 2: No crop-specific data available for my region

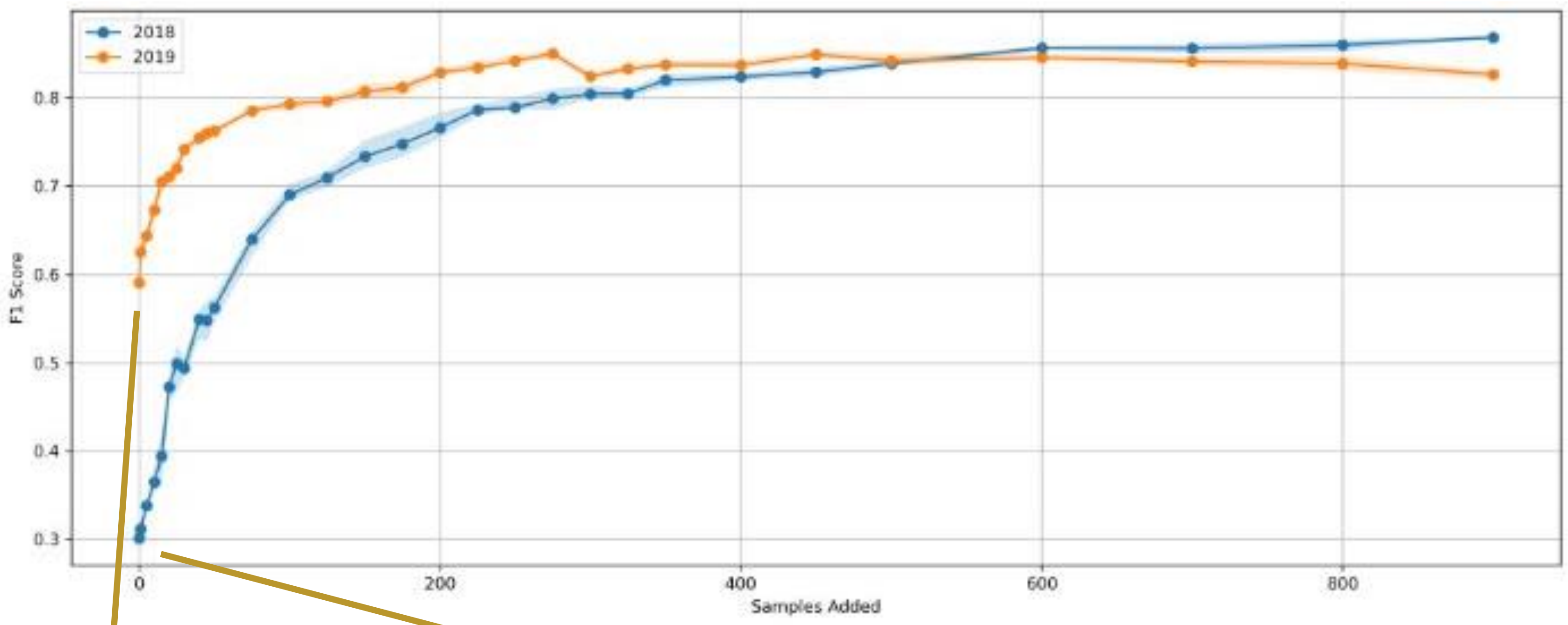
Cassava classification in Mozambique



Baseline: No cassava data available for Mozambique, model trained with other crops from Mozambique and cassava data from other regions

Case 3: Benefit of year-specific reference data

Model performance as more year-specific data gets added (Argentina)



- Models for both 2018 and 2019 show consistent improvement with increasing sample injections.
- The 2019 model starts at a higher baseline (~0.7 F1) than the 2018 model (~0.5 F1) due to the availability of multi-year data (2017 + 2018).
- The 2019 model approaches high performance (~0.8 F1) much faster with fewer injections compared to the 2018 model, highlighting the advantage of leveraging historical data for future year predictions.
- Incremental injections from the target year lead to consistent improvements.

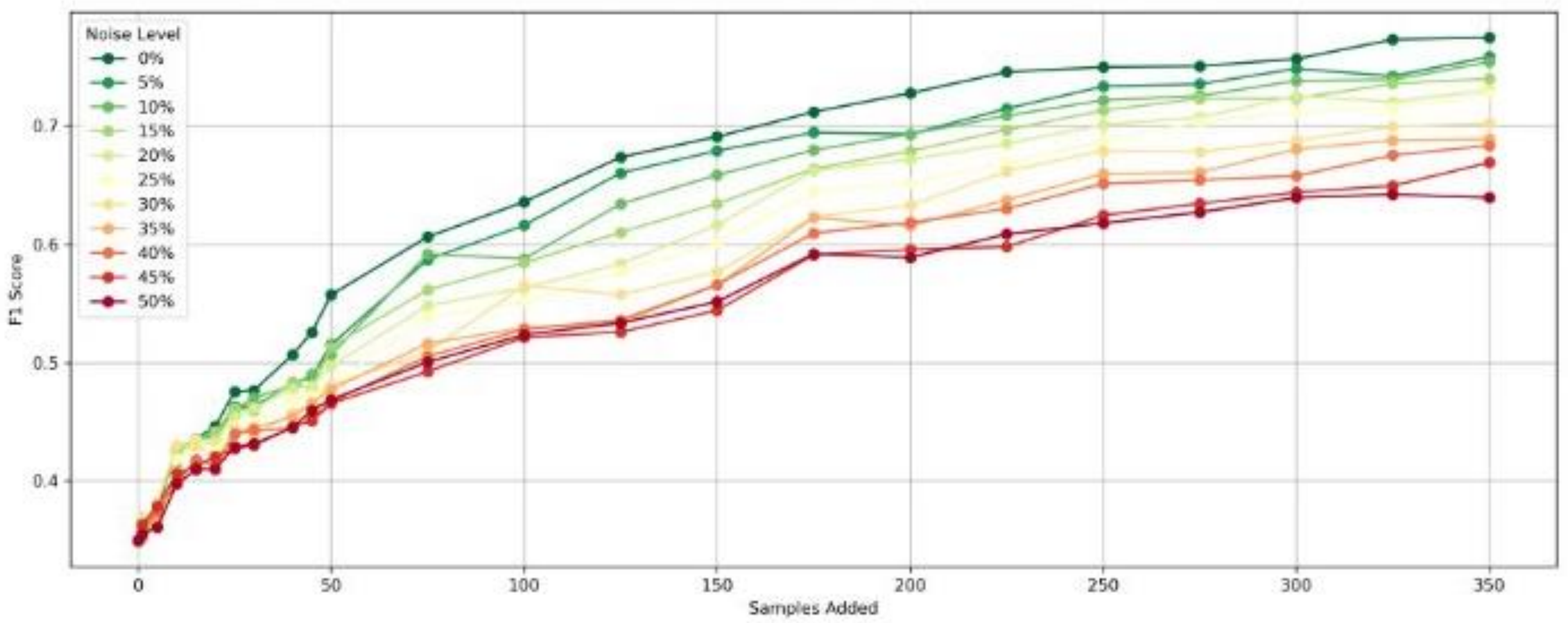
Baseline: Model trained using data from 2017

Baseline: Model trained using data from 2017 + 2018

WorldCereal

Case 4: Impact of data quality on model performance?

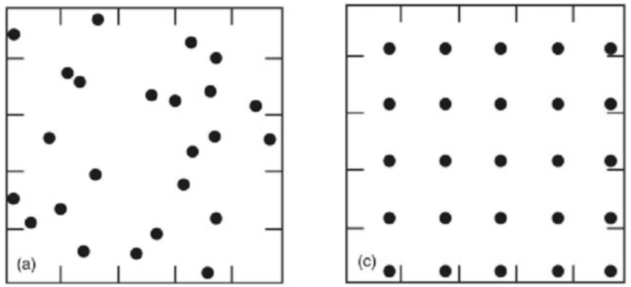
Improvement in model performance in Argentina upon adding data of different quality



- Even small amounts of noise in training labels degrade performance, with the impact increasing as more data is injected.
- Small injection sizes tolerate noise better, as the model depends more on cleaner patterns from the base dataset.
- As injection size grows, ensuring data quality becomes increasingly important to avoid propagating noisy patterns.
- **For optimal performance, prioritize clean, high-quality training data, especially when working with large datasets.**

WorldCereal

So, how much reference data do I need?

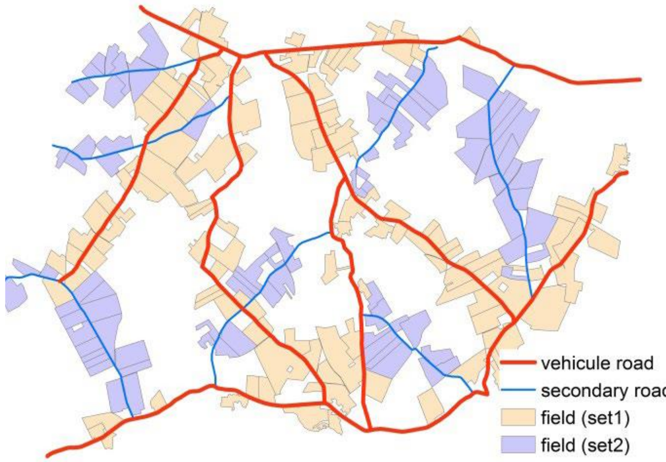


Model training

- The total area of interest should first be **stratified** into different agro-ecological zones. Once achieved, **50-100 samples** for each main crop and 20-30 samples for each minor crop should be collected in each of the delineated zones.

- **1 observation/sq. km – 1 observation/10 sq. km** for major crop types, depending on landscape complexity and variability

- Ideally, random or systematic sampling design. Practically -> **windshield surveys**!



Statistical validation of generated maps

Requires totally different approach!
Define regular sampling scheme based on the product you want to validate…
Will be discussed in depth during the third WorldCereal MOOC

Key Takeaways

1. Invest in local data for your crops of interest
Adding 10-50 relevant samples already makes a huge difference! Larger batches of 100-500 are needed to optimize performance.

2. Make sure to leverage existing datasets from other regions/years in your applications
Models trained with a diverse set of data provide a strong baseline for further local improvements. Consult and use the public datasets in the WorldCereal RDM! Check the performance of your baseline model before deciding on the number of samples to be collected.

3. Prioritize data quality
Clean and high-quality data is critical, especially for large-scale data injections. Rather have few good quality samples, than lots of mediocre quality samples.

WorldCereal

Further reading

Several studies cover the same topic of importance of adding relevant training data. These studies collectively reinforce and echo our analysis in concluding the benefits of incorporating high-quality, localized, and crop-specific data.

1. Toward Sustainability: Trade-Off Between Data Quality and Quantity in Crop Pest Recognition
2. Meta-Learning for Few-Shot Land Cover Classification
3. A Bayesian-inspired, deep learning-based, semi-supervised domain adaptation technique for land cover mapping
4. On the Generalizability of Foundation Models for Crop Type Mapping
5. MTP: Advancing Remote Sensing Foundation Model via Multitask Pretraining
6. Few-Shot Learning for Crop Mapping from Satellite Image Time Series
7. EUROCROPSML: A Time Series Benchmark Dataset For Few-Shot Crop Type Classification
8. Generalized few-shot learning for crop hyperspectral image precise classification
9. A survey of few-shot learning in smart agriculture: developments, applications, and challenges

WorldCereal

WorldCereal

THANK YOU

Interesting links:

About ref data → https://esa-worldcereal.org/en/reference-data

RMD UI → https://rdm.esa-worldcereal.org/

Documentation → https://worldcereal.github.io/worldcereal-documentation/rdm/overview.html

Questions? → WorldCereal Forum MOOC I

Subscribe to our mailing list