WorldCereal

esa

# WorldCereal MOOC I

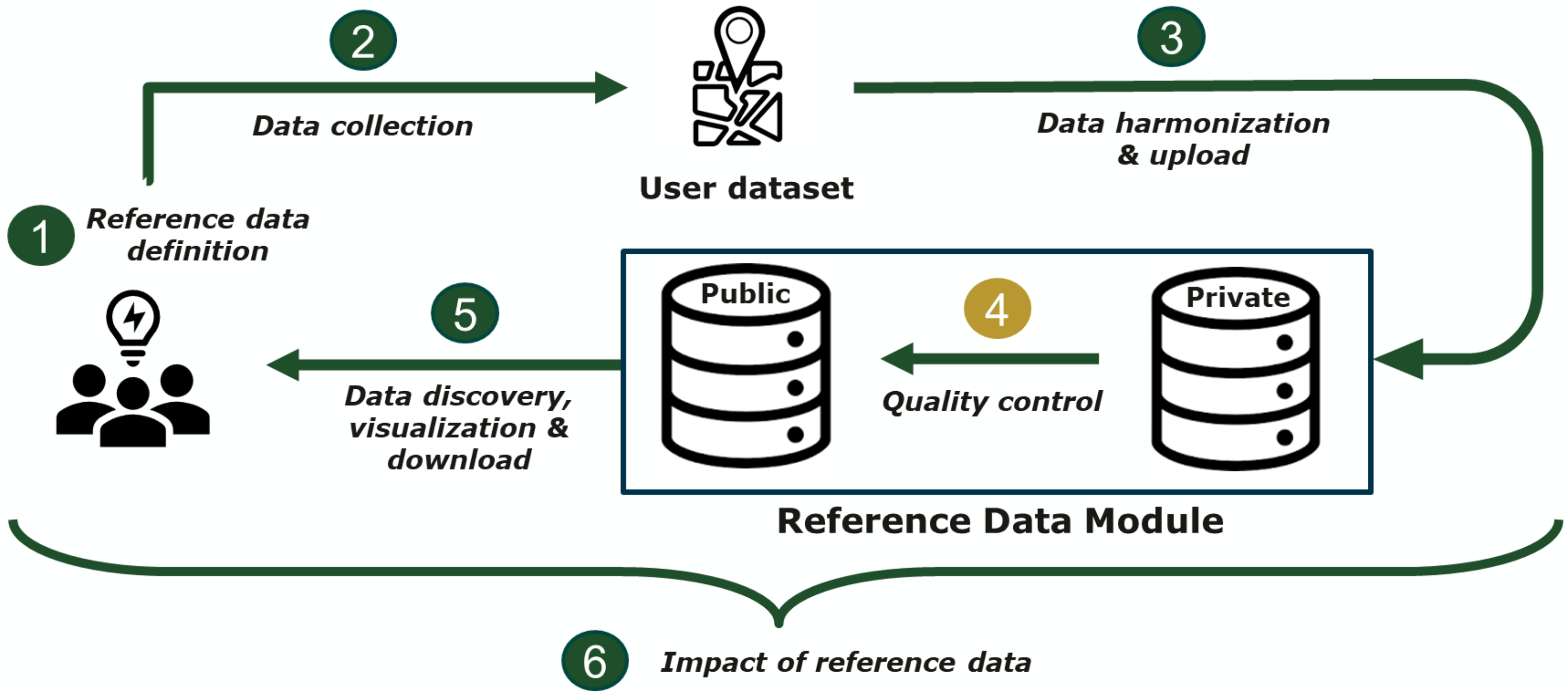
## Quality assessment of reference data

Arun Pratihast (WENR)

vito · WAGENINGEN UNIVERSITY & RESEARCH · IIASA · UNIVERSITÉ DE STRASBOURG · UNIVERSITAT DE VALÈNCIA · gisat

→ THE EUROPEAN SPACE AGENCY

# MOOC I: Outline



1 Reference data definition

2 Data collection

User dataset

3 Data harmonization & upload

Public

4 Quality control

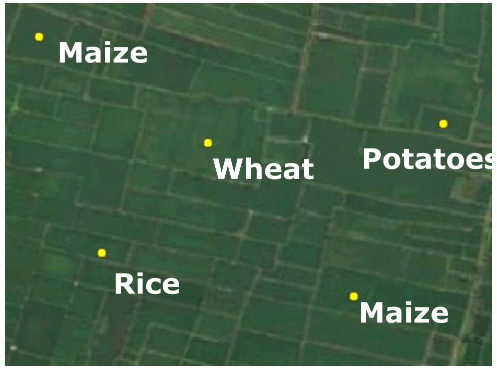
Private

Reference Data Module

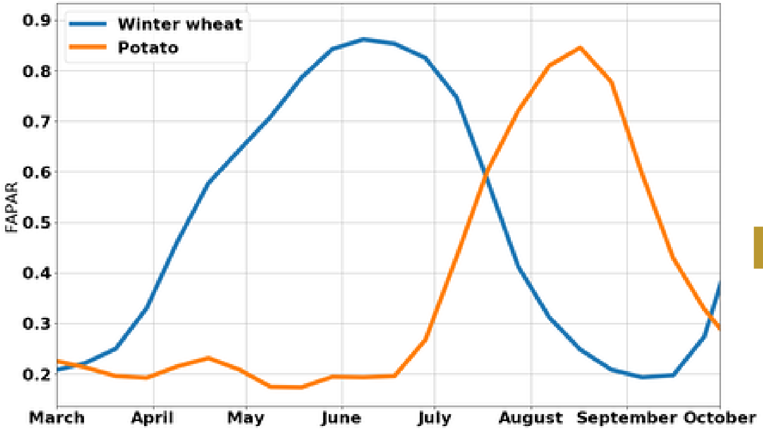
5 Data discovery, visualization & download

6 Impact of reference data

# Mapping crops from space

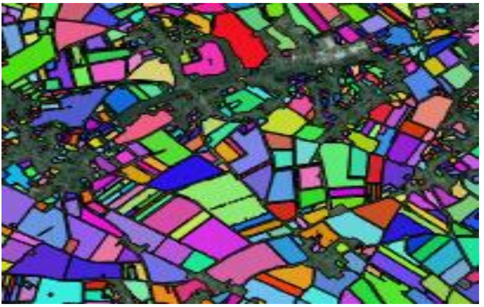


Reference data



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

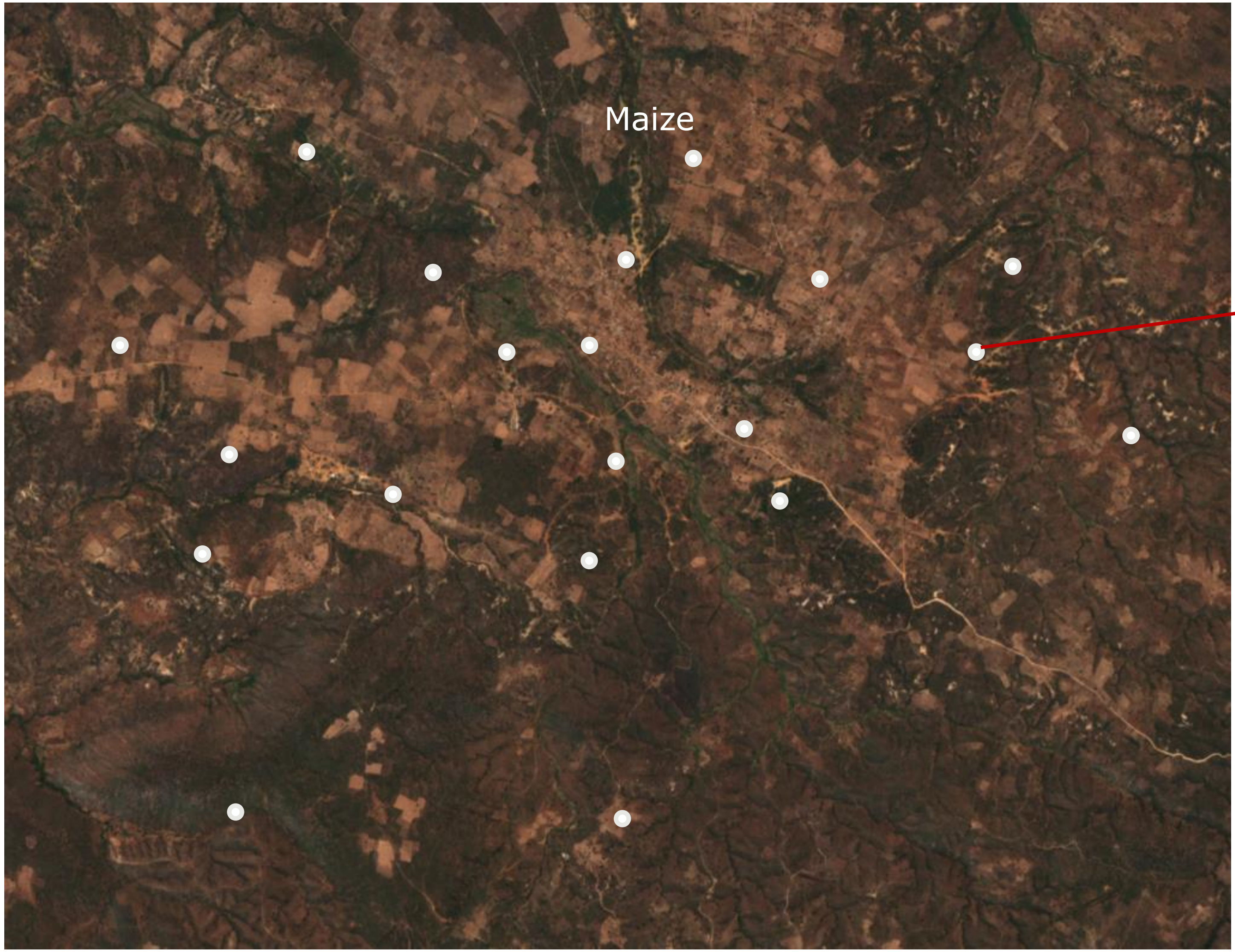
Crop identification model

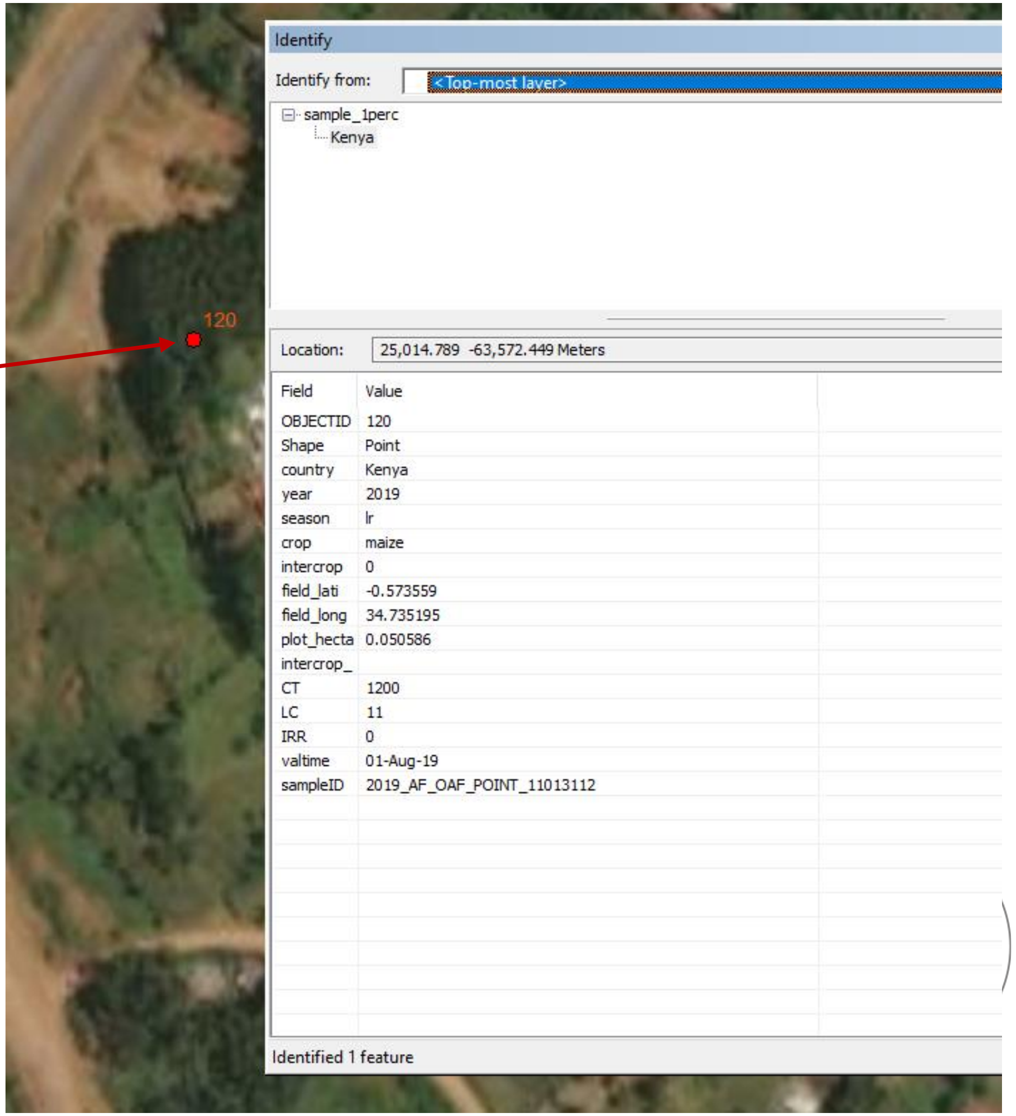


Crop type map

WorldCereal

# Reference data





Identify

Identify from: <Top-most layer>

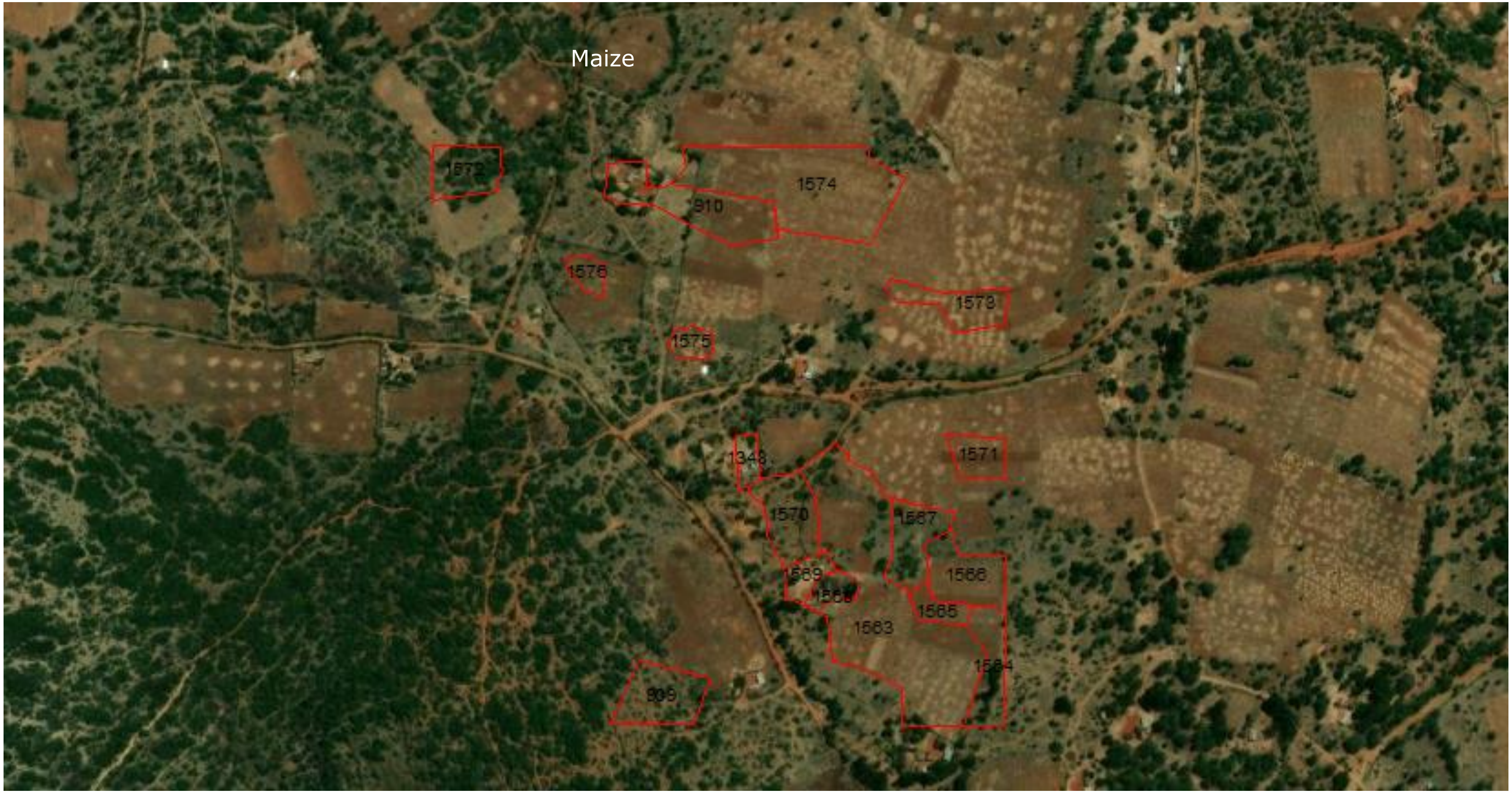
sample_1perc
Kenya

Location: 25,014.789 -63,572.449 Meters

| Field | Value |
|---|---|
| OBJECTID | 120 |
| Shape | Point |
| country | Kenya |
| year | 2019 |
| season | lr |
| crop | maize |
| intercrop | 0 |
| field_lati | -0.573559 |
| field_long | 34.735195 |
| plot_hecta | 0.050586 |
| intercrop_ | |
| CT | 1200 |
| LC | 11 |
| IRR | 0 |
| valtime | 01-Aug-19 |
| sampleID | 2019_AF_OAF_POINT_11013112 |

Identified 1 feature

# Reference data



# Factors affecting reference data collection

### Access and Logistics

Remote locations and infrastructure challenges hinder data collection.

### Field Boundary Delineation

Complex shapes and unclear boundaries lead to mapping inaccuracies.

### Human and Equipment Errors

Training and tool issues reduce data quality.

### GPS Accuracy

Poor signals and terrain interfere with precise geolocation.

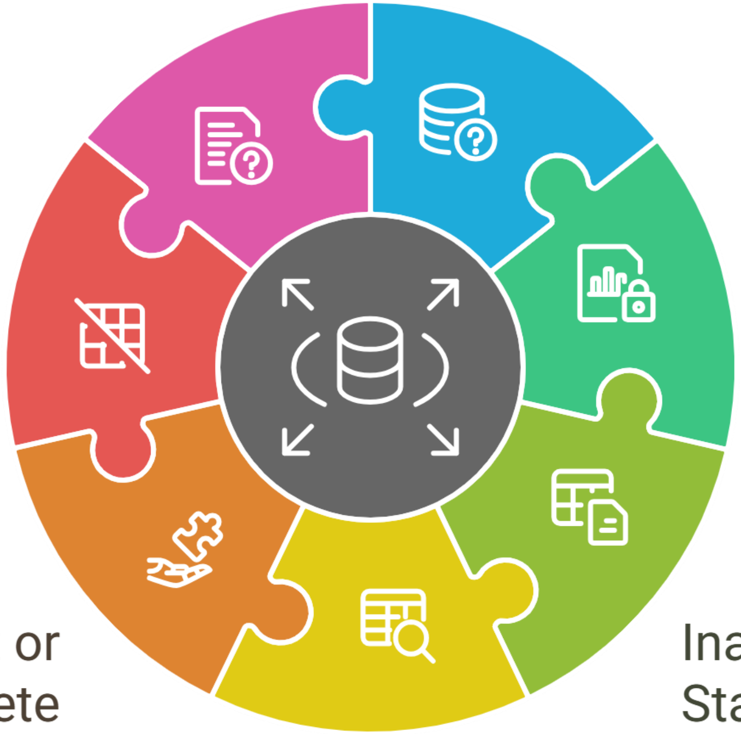
### Crop Calendar Mismatches

Timing variations result in surveys at wrong growth stages.

### Weather Variability

Environmental factors disrupt surveys and alter crop conditions.

# Factors affecting re-use of reference data



Unclear (Re)Use Data Policies

Unknown Data Collection Protocols

Limited Data Accessibility and Formats

Inadequate Data Standardization

Unknown Data Quality

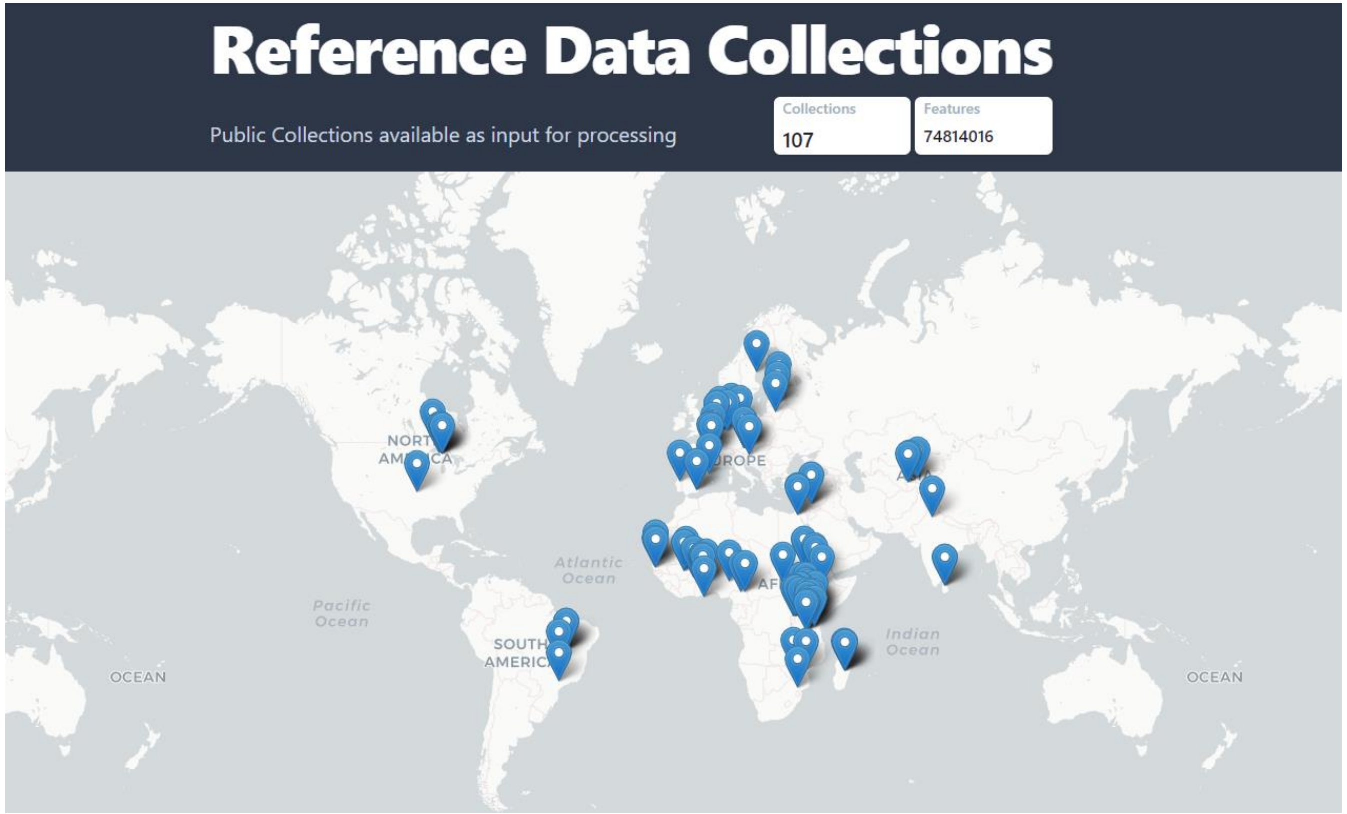
Inconsistent or Incomplete Metadata

Limited Coverage

WorldCereal

# Factors affecting reference data quality

- At WorldCereal, we collected more than 100 datasets, comprising approximately 75 million observations
- Each dataset follows its own data collection protocol, features distinct attributes, and serves different purposes
- Challenge was on how to develop a generic framework to evaluate the quality of the datasets?



## Reference Data Collections

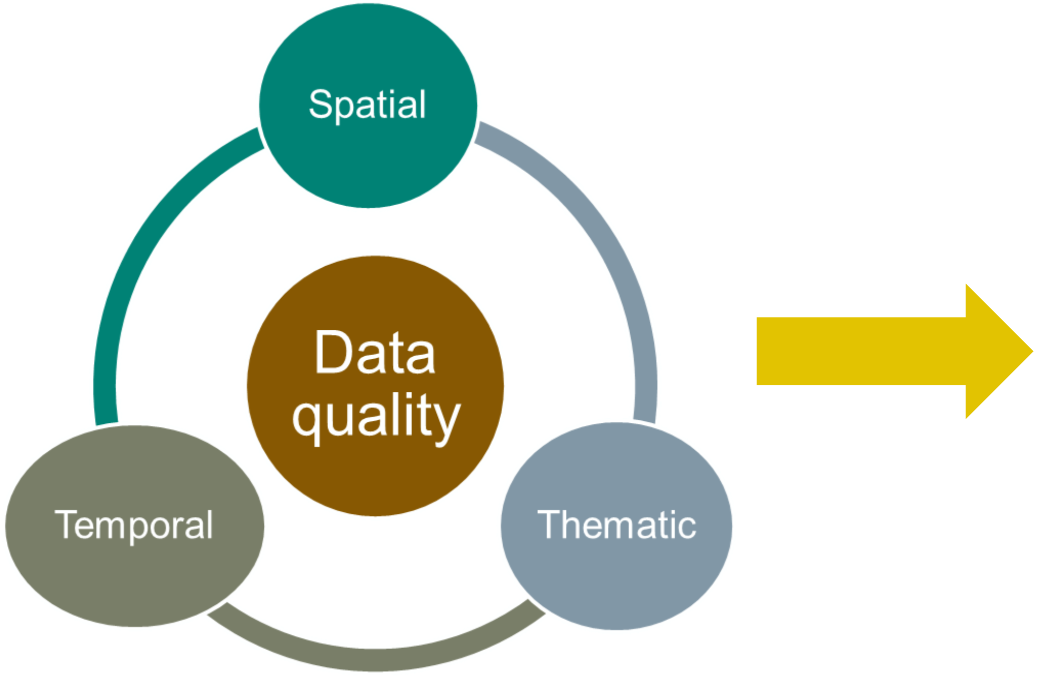
Public Collections available as input for processing

| Collections | Features |
|---|---|
| 107 | 74814016 |

WorldCereal

# WorldCereal reference data: quality assessment

- First version of a generic framework for evaluating the quality and assigning single confidence score to each dataset
- Confidence score reflects the fitness for use as reference data for training Earth Observation based crop classification algorithms



```
Step 1: IF No geo-locations THEN
        Data set rejected
Step 2: IF Date ranges not between 2017 till date THEN
        Data set rejected
Step 3: IF No WorldCereal cropland and/or crop type THEN
        Data set rejected
Step 4: ELSE
        $\text{Average confidence score} = \frac{\sum_1^n Q_i * W_i}{100}$
Where: Q = Quality score (ranges from 0-100); W= weight factor per accuracy category
and i = accuracy category ranges from 1 to n.
```

# Data quality dimensions

## Geometry / Spatial Accuracy

- Precision of feature positions in a spatial reference system
- Compared to reference data 'true’ position
- **WorldCereal**: Evaluated for vector datasets (GPS errors, spatial context) and raster datasets (spatial resolution)

## Temporal Accuracy

- Accuracy of time components (acquisition date, estimated dates, etc.)
- **WorldCereal**: Linked to specific dates from observations or satellite imagery

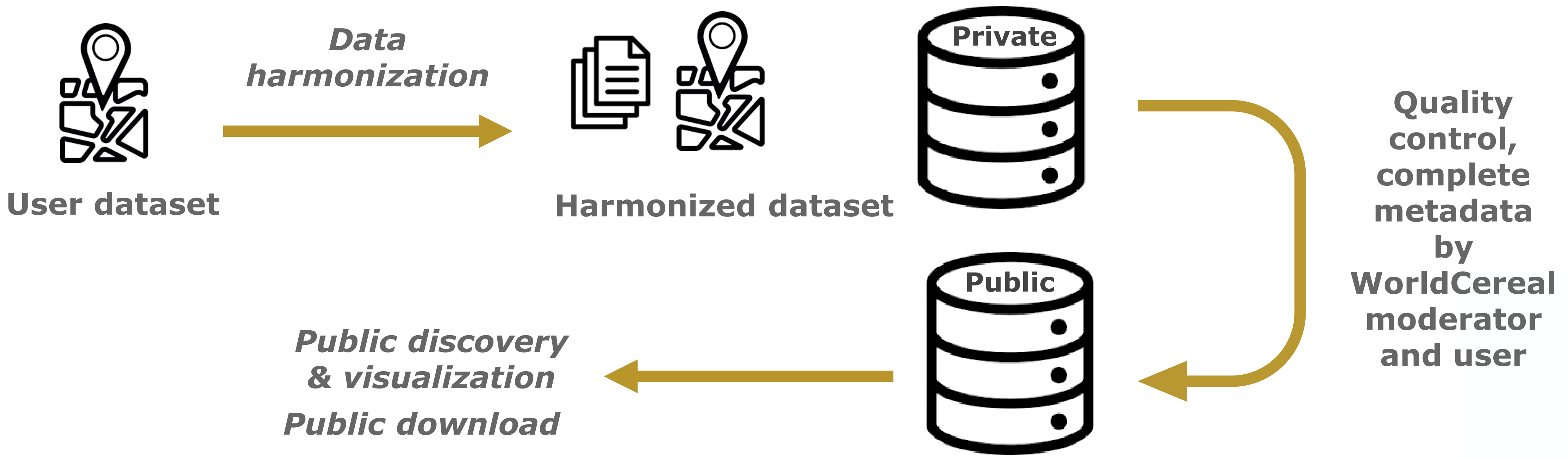
## Thematic Accuracy

- Accuracy of thematic tags (land cover, crop type)
- Related to validation methods and classification accuracy
- **WorldCereal**: Assessed based on validation methods, user confidence, and automated classification

# Data review and quality assessment process

- Public: share your reference data with public
- Private: only you can use the data

→ **Additional metadata quality control !**



*Data harmonization*

**User dataset**

**Harmonized dataset**

**Private**

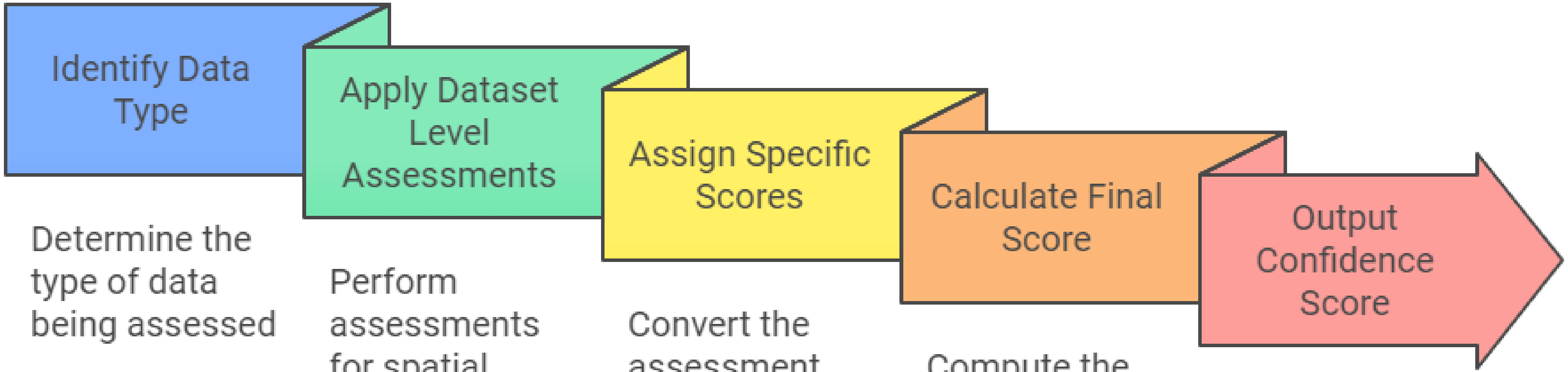
**Public**

**Quality control, complete metadata by WorldCereal moderator and user**

*Public discovery & visualization*

*Public download*

# Confidence score calculation process



**Identify Data Type**

Determine the type of data being assessed

**Apply Dataset Level Assessments**

Perform assessments for spatial, temporal, and thematic accuracy

**Assign Specific Scores**

Convert the assessment results into confidence scores for each category

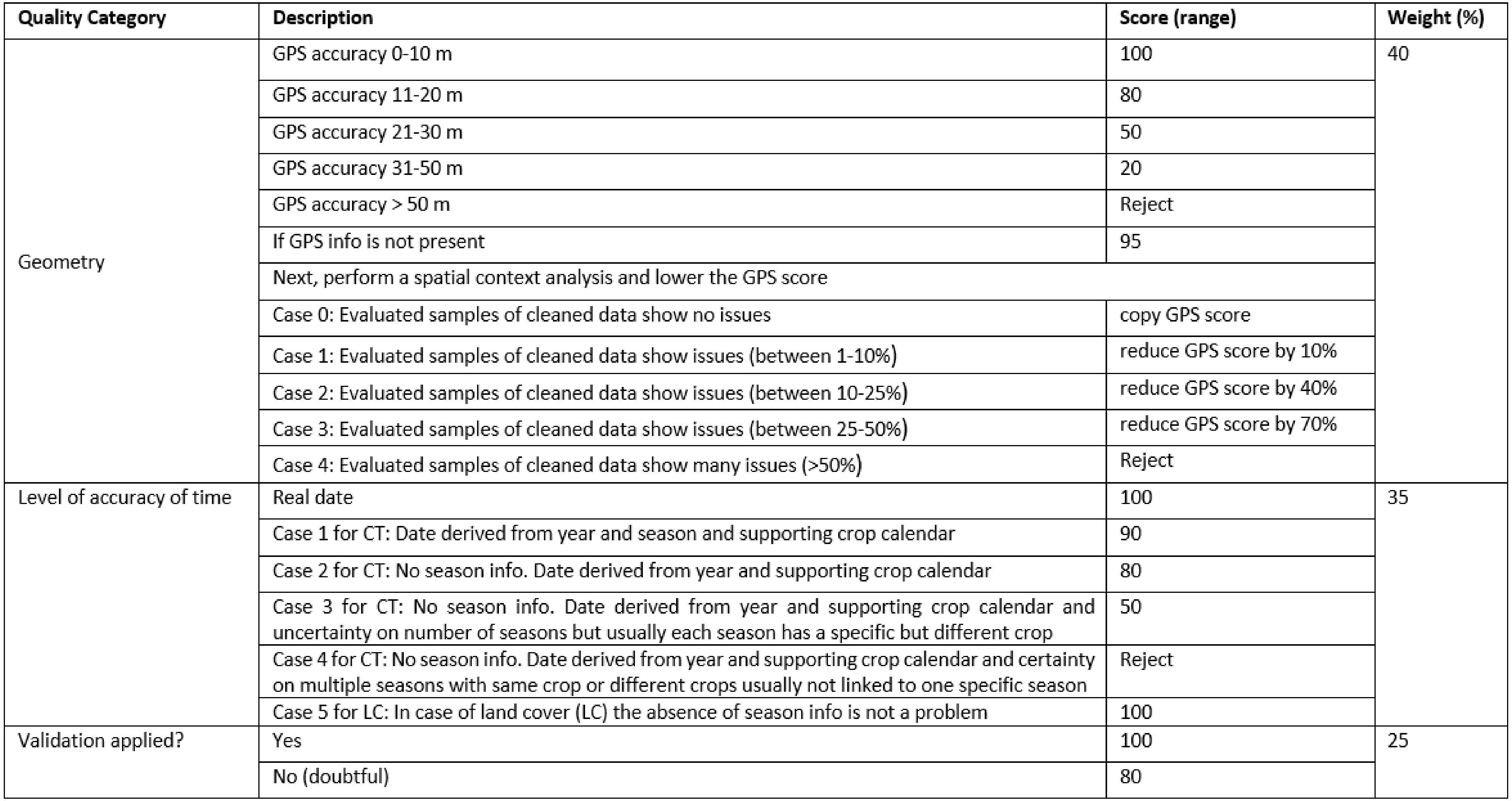
**Calculate Final Score**

Compute the overall confidence score by averaging the individual scores

**Output Confidence Score**

Return the final calculated confidence score

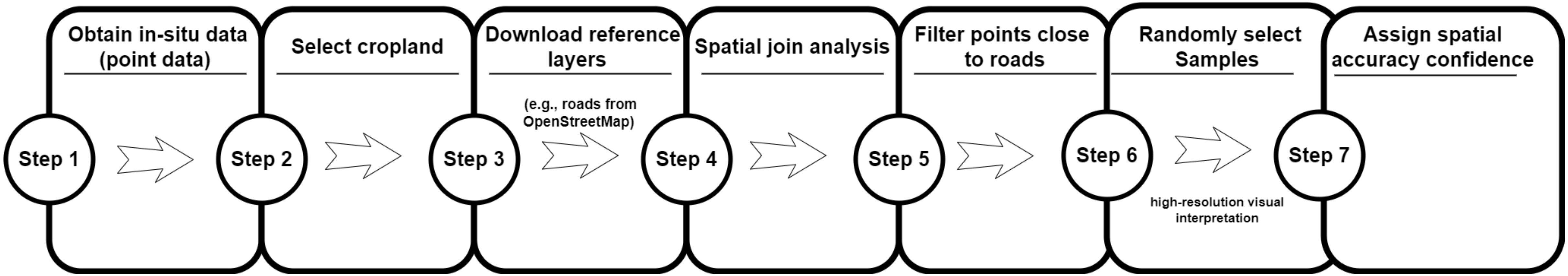
# Data quality assessment process



| Quality Category | Description | Score (range) | Weight (%) |
|---|---|---|---|
| Geometry | GPS accuracy 0-10 m | 100 | 40 |
| | GPS accuracy 11-20 m | 80 | |
| | GPS accuracy 21-30 m | 50 | |
| | GPS accuracy 31-50 m | 20 | |
| | GPS accuracy > 50 m | Reject | |
| | If GPS info is not present | 95 | |
| | Next, perform a spatial context analysis and lower the GPS score | | |
| | Case 0: Evaluated samples of cleaned data show no issues | copy GPS score | |
| | Case 1: Evaluated samples of cleaned data show issues (between 1-10%) | reduce GPS score by 10% | |
| | Case 2: Evaluated samples of cleaned data show issues (between 10-25%) | reduce GPS score by 40% | |
| | Case 3: Evaluated samples of cleaned data show issues (between 25-50%) | reduce GPS score by 70% | |
| | Case 4: Evaluated samples of cleaned data show many issues (>50%) | Reject | |
| Level of accuracy of time | Real date | 100 | 35 |
| | Case 1 for CT: Date derived from year and season and supporting crop calendar | 90 | |
| | Case 2 for CT: No season info. Date derived from year and supporting crop calendar | 80 | |
| | Case 3 for CT: No season info. Date derived from year and supporting crop calendar and uncertainty on number of seasons but usually each season has a specific but different crop | 50 | |
| | Case 4 for CT: No season info. Date derived from year and supporting crop calendar and certainty on multiple seasons with same crop or different crops usually not linked to one specific season | Reject | |
| | Case 5 for LC: In case of land cover (LC) the absence of season info is not a problem | 100 | |
| Validation applied? | Yes | 100 | 25 |
| | No (doubtful) | 80 | |

WorldCereal

# Spatial accuracy assessment



Step 1 – Obtain in-situ data (point data)

Step 2 – Select cropland

Step 3 – Download reference layers

(e.g., roads from OpenStreetMap)

Step 4 – Spatial join analysis

Step 5 – Filter points close to roads

Step 6 – Randomly select Samples

high-resolution visual interpretation

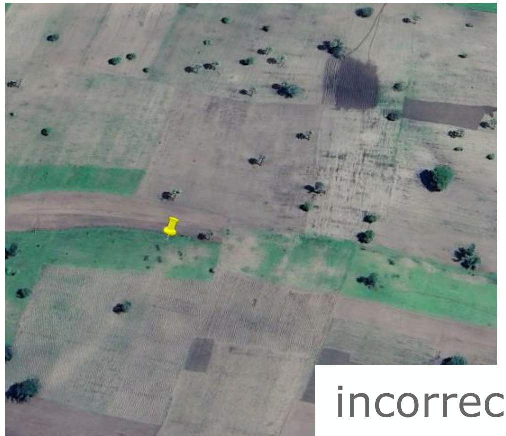
Step 7 – Assign spatial accuracy confidence

| Features Range | Percentage for Visual Inspection |
|---|---|
| 0 - 20 | 50% |
| 21 - 50 | 20% |
| 51 - 100 | 10% |
| 101 - 200 | 7.5% |
| 201 - 500 | 5% |
| 501 - 1000 | 3% |
| 1001 - 5000 | 2% |
| 5001 - 20000 | 1% |
| 20001 - 50000 | 0.5% |

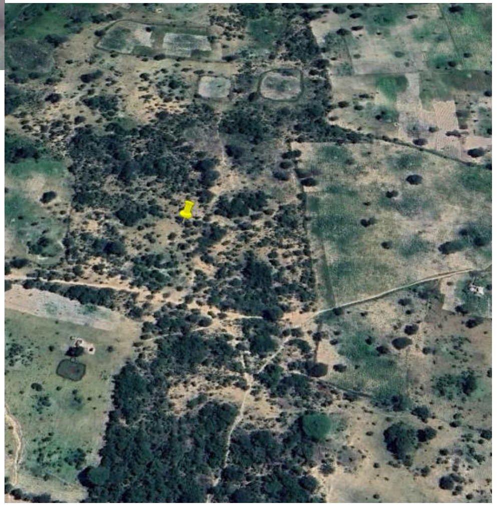
WorldCereal

# Spatial accuracy assessment

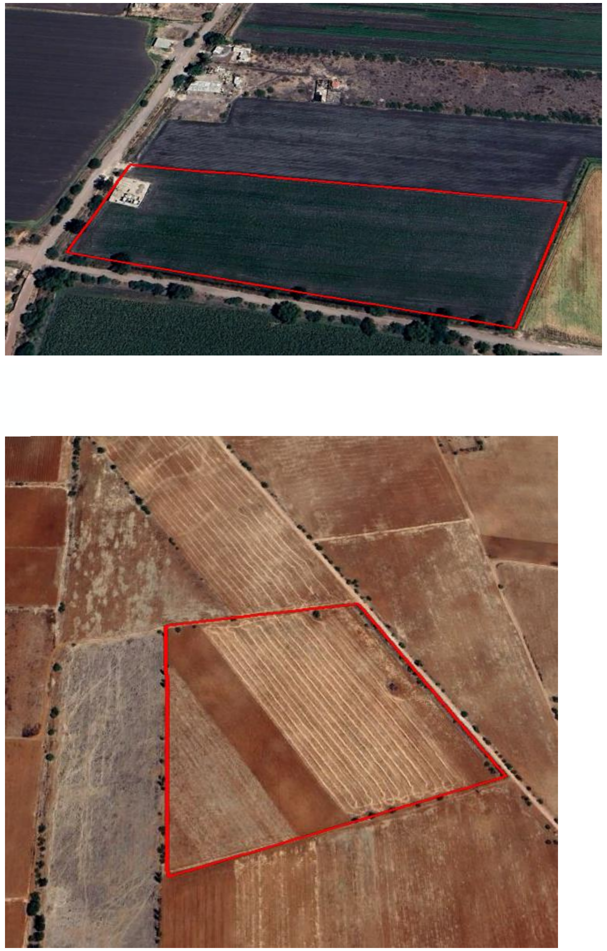
correct cropland point



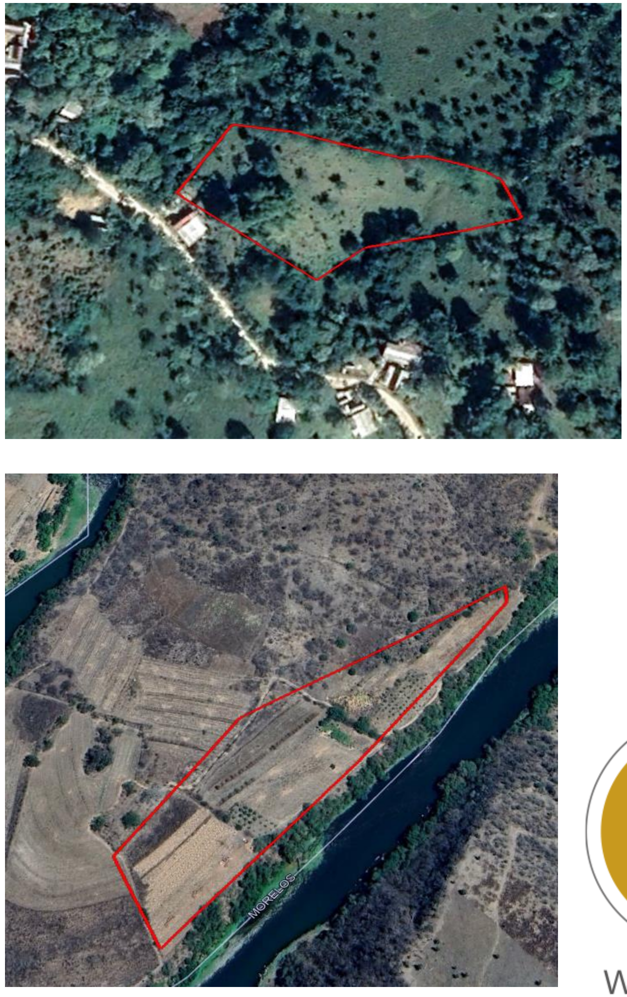
incorrect cropland point



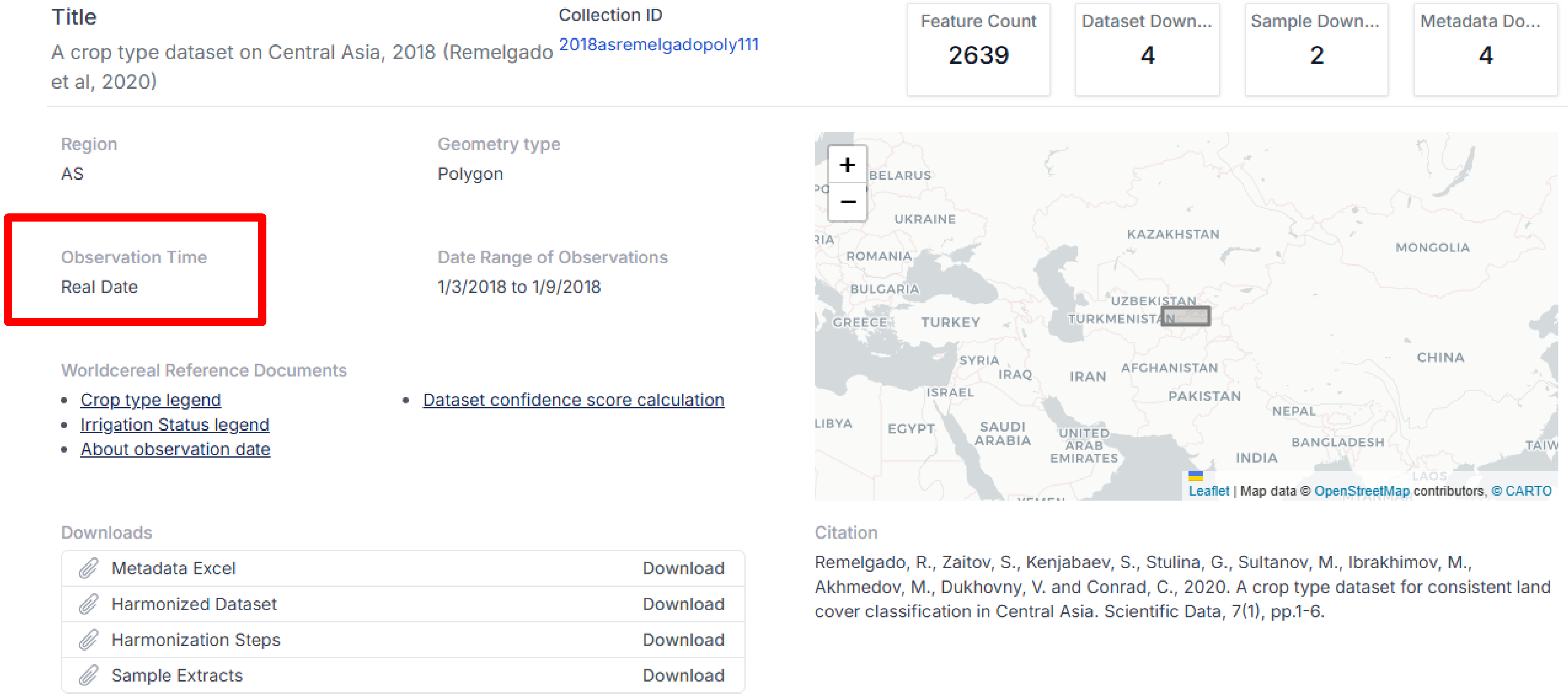
correct cropland polygons



incorrect cropland polygons



# Example metadata - Remelgado et al., 2020



Title

A crop type dataset on Central Asia, 2018 (Remelgado et al, 2020)

Collection ID

2018asremelgadopoly111

| Feature Count | Dataset Down... | Sample Down... | Metadata Do... |
|---|---|---|---|
| 2639 | 4 | 2 | 4 |

Region

AS

Geometry type

Polygon

Observation Time

Real Date

Date Range of Observations

1/3/2018 to 1/9/2018

Worldcereal Reference Documents

- Crop type legend
- Irrigation Status legend
- About observation date
- Dataset confidence score calculation

Downloads

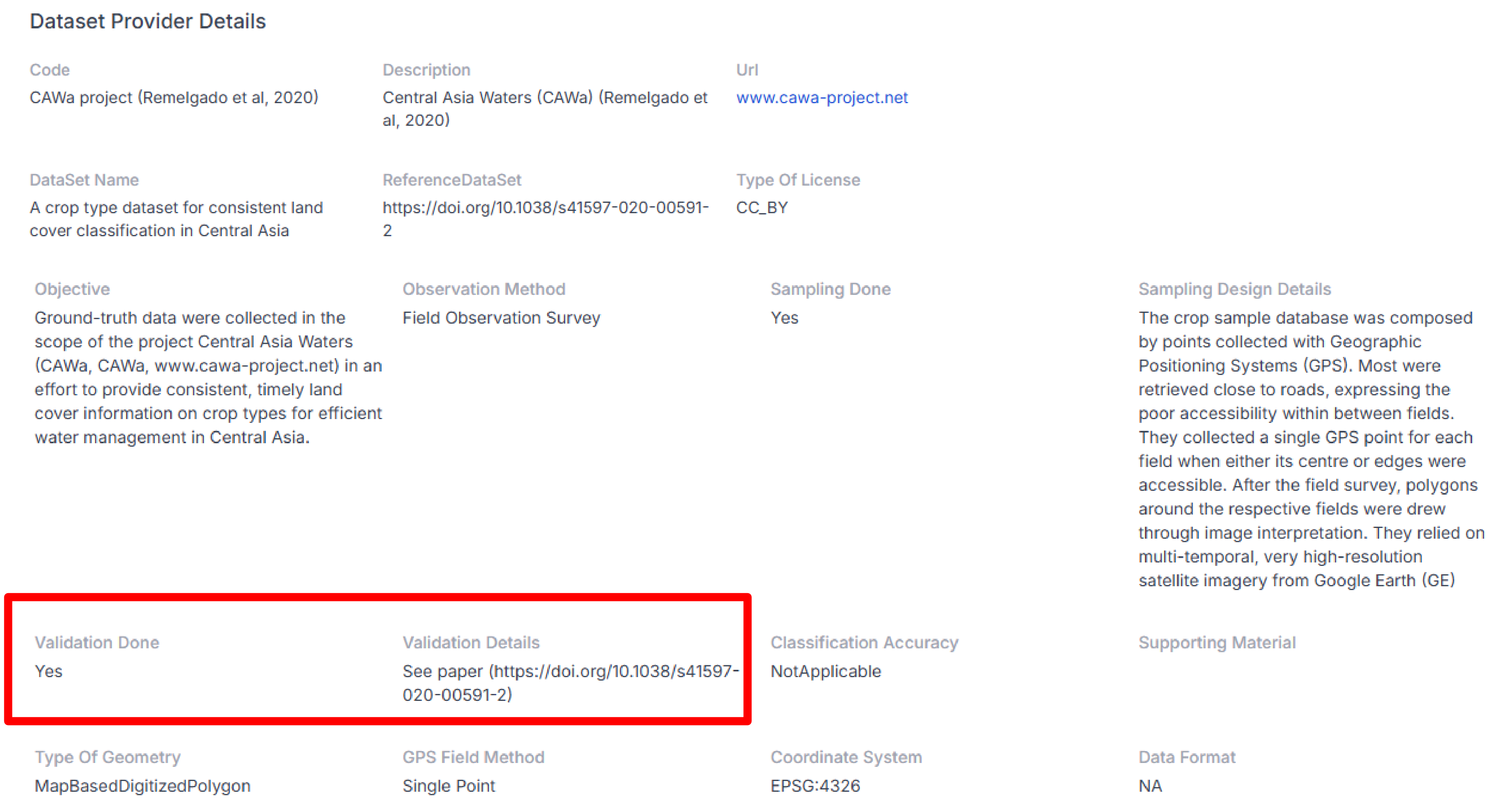
| | |
|---|---|
| Metadata Excel | Download |
| Harmonized Dataset | Download |
| Harmonization Steps | Download |
| Sample Extracts | Download |

Citation

Remelgado, R., Zaitov, S., Kenjabaev, S., Stulina, G., Sultanov, M., Ibrakhimov, M., Akhmedov, M., Dukhovny, V. and Conrad, C., 2020. A crop type dataset for consistent land cover classification in Central Asia. Scientific Data, 7(1), pp.1-6.

WorldCereal

# Example metadata - Remelgado et al., 2020



### Dataset Provider Details

| Code | Description | Url |
|---|---|---|
| CAWa project (Remelgado et al, 2020) | Central Asia Waters (CAWa) (Remelgado et al, 2020) | www.cawa-project.net |

| DataSet Name | ReferenceDataSet | Type Of License |
|---|---|---|
| A crop type dataset for consistent land cover classification in Central Asia | https://doi.org/10.1038/s41597-020-00591-2 | CC_BY |

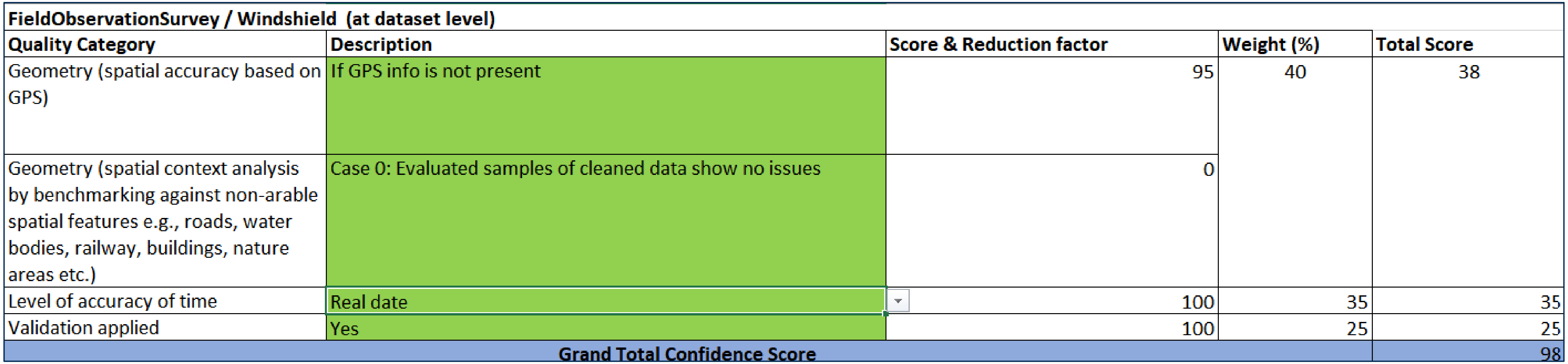
| Objective | Observation Method | Sampling Done | Sampling Design Details |
|---|---|---|---|
| Ground-truth data were collected in the scope of the project Central Asia Waters (CAWa, CAWa, www.cawa-project.net) in an effort to provide consistent, timely land cover information on crop types for efficient water management in Central Asia. | Field Observation Survey | Yes | The crop sample database was composed by points collected with Geographic Positioning Systems (GPS). Most were retrieved close to roads, expressing the poor accessibility within between fields. They collected a single GPS point for each field when either its centre or edges were accessible. After the field survey, polygons around the respective fields were drew through image interpretation. They relied on multi-temporal, very high-resolution satellite imagery from Google Earth (GE) |

| Validation Done | Validation Details | Classification Accuracy | Supporting Material |
|---|---|---|---|
| Yes | See paper (https://doi.org/10.1038/s41597-020-00591-2) | NotApplicable | |

| Type Of Geometry | GPS Field Method | Coordinate System | Data Format |
|---|---|---|---|
| MapBasedDigitizedPolygon | Single Point | EPSG:4326 | NA |

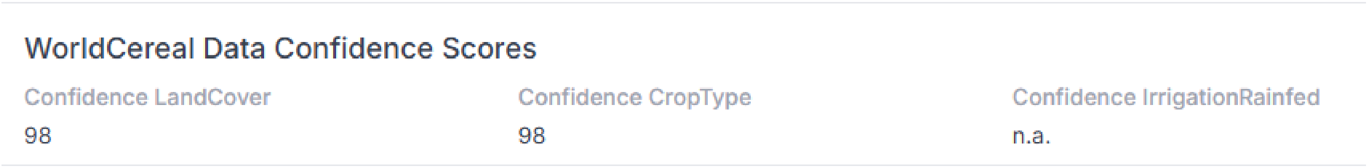
WorldCereal

# Example metadata - Remelgado et al., 2020



**FieldObservationSurvey / Windshield (at dataset level)**

| Quality Category | Description | Score & Reduction factor | Weight (%) | Total Score |
|---|---|---|---|---|
| Geometry (spatial accuracy based on GPS) | If GPS info is not present | 95 | 40 | 38 |
| Geometry (spatial context analysis by benchmarking against non-arable spatial features e.g., roads, water bodies, railway, buildings, nature areas etc.) | Case 0: Evaluated samples of cleaned data show no issues | 0 | | |
| Level of accuracy of time | Real date | 100 | 35 | 35 |
| Validation applied | Yes | 100 | 25 | 25 |
| **Grand Total Confidence Score** | | | | 98 |



WorldCereal Data Confidence Scores

| Confidence LandCover | Confidence CropType | Confidence IrrigationRainfed |
|---|---|---|
| 98 | 98 | n.a. |

# In summary

- We developed a set of rules to assess the spatial, temporal, and thematic quality of each dataset summarized in one single confidence score
- Visual interpretation challenges:
  - Ambiguities in high-resolution imagery
  - Variability among analysts
  - Labor-Intensive
- This quality assessment is typically performed by the WorldCereal moderator when users decide to make their data publicly available through the WorldCereal Reference Data Module (RDM)

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

esa · vito · gisat · IIASA · UNIVERSITÉ DE STRASBOURG · VNIVERSITAT DE VALÈNCIA · WAGENINGEN UNIVERSITY & RESEARCH