Scheme to calculate confidence (quality) scores

(WorldCereal Reference Data)

Document Ref:	WorldCereal_ConfidenceScoreCalculations_v1_1
Version:	1.1
Creation Date:	2024-06-16
Last Modified:	2024-09-18

1 Contents

2 Introduction

To decide on the fitness for use of a reference dataset for training of classification algorithms or the validation of the final products in WorldCereal phase1, we developed a generic scheme to calculate confidence scores. The scheme differentiates 4 different data types and includes the spatial, temporal, and thematic accuracy since these are essential aspects of crop mapping. Additionally, we refined these rules in WorldCereal Phase 2.

The spatial accuracy refers to the accuracy of the position of spatial features within a spatial reference system and is usually assessed by comparing the position of features with their counterparts in reference data, which are considered to represent the 'true' position. In WorldCereal, we assessed the geometry of vector and raster datasets. The geometry of vector datasets refers to GPS recording errors and in addition to the spatial context (e.g. was the observation really observed within the field or from an adjacent road). The spatial resolution of a raster refers to the size of grid/pixel in a raster dataset.

The temporal accuracy refers to the accuracy of assessing the validity time. The data set may have an actual observation date. Alternatively, the validity time is derived from the observed calendar year (and season) and governing local crop calendars. This affects accuracy especially if multiple cropping seasons exist and crop calendars have wide planting and harvesting windows and/or representing large regions characterized by complex cropping systems. Concerning crowdsourcing or expert campaigns (Classification or Validation by crowd or expert) validity time is preferably based on the satellite imagery time and not derived from the submission date e.g. the date that the expert or crowd submitted their assessment.

Finally, the thematic accuracy refers to the accuracy of the thematic labels associated with the datasets such as presence of LC, CT, and IRR. This can be linked to evidence on validation and quality control (Field Observation), the user confidence (Classification or Validation by crowd or expert) and classification accuracy (Automated Classification).

Below the weight factors per accuracy category (see for justification the detailed tables below).

	Geometry/Spatial accuracy	Level of accuracy of time	Thematic accuracy
Field Observation	40	35	25
Classification or Validation by crowd or expert	40	25	35
Automated Classification	35	25	40
Formal Declaration	40	30	30

The details of the calculation measures are listed below:

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

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.

We assess confidence at dataset level. In the case of type “Classification or Validation by crowd or expert” user confidence first needs to be summarized from sample-level to an “average” data set level.

3 Step 1: assess spatial accuracy for Field Observation (FO)

Step 1: Protocol to assess spatial accuracy for Field Observation (FO) (at dataset level). This can be skipped for other data types (Classification or Validation by crowd or expert, Automated Classification, Formal Declaration)

If Field Observation (FO) is point data, proceed with the following steps:

Step 1: Select the crop land labels

Step 2: Find the bounding box of the datasets

Step 3: Download OSM layers (e.g., roads, water bodies, railway, buildings, nature areas, etc.)

Step 4: Calculate spatial distances between the crop land label points and roads

Step 5: Remove points that are less than 20 meters from roads

Step 6: Select samples for visual inspection based on the following rule:

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%

Step 7: Identify the number of suspicious cases (close to or overlapping infrastructure, very small field (<20m wide/long), obviously wrong cover e.g. expect maize but mainly trees etc) via visual inspection of samples (cropland: yes or no)

Step 8: Calculate the percentage of suspicious cases (=no) of the total number of cleaned cropland labels

Percentage	Penalty
0 - 1%	No penalty
1 - 10%	Lower score by 10%
10 - 25%	Lower score by 40%
25 – 50%	Lower score by 70%
>50%	Lower score by 100%

If Field Observation (FO) is polygon data, proceed with the following steps:

Step 1: Select the crop land labels

Step 2: Find the bounding box of the datasets

Step 3: Download the OSM layers (e.g., roads, water bodies, railway, buildings, nature areas, etc.).

Step 4: Determine the number of intersection cases in the cropland file by overlaying with roads and generate samples for visual interpretation

Step 5: Remove polygons that intersect roads

Step 6: Select samples for visual inspection same as point data (see above)

Step 7: Identify the number of suspicious cases (close to or overlapping infrastructure, very small field (<20m wide/long), obviously wrong cover e.g. expect maize but mainly trees etc) via visual inspection of samples (cropland: yes or no)

Step 8: Calculate the percentage of suspicious cases (no) of the total number of cropland labels

Note Python scripts have been prepared to clean point and polygon data using OSM-based road data, save distance to road (points) in separate attribute, select random samples for visual inspections, calculate the percentage of suspicious cases and the percentage of intersection with roads (polygons).

4 Step 2: determine the confidence scores

Step 2: Determine the confidence scores

The rules to calculate the confidence score is explained on the following pages.

These rules have been implemented in the Excel tool "WorldCereal_DataConfidenceScore_Calculator_v3_0". As the rules slightly differ for crop type and landcover the moderator, responsible for determining the confidence scores, has to apply the tool twice for crop type and land cover.

Details calculations for Field Observation (FO) (at dataset level)

Quality Category	Description	Score (range)	Weight (%)	Justification
Geometry	GPS accuracy 0-10 m	100	40	If the GPS error is reported in data sets then start from there. In case no information on GPS is given we apply a small penalty. Next, we run additional checks to detect suspicious cases for which cropland intersects road infrastructure. See above for separate protocol on spatial accuracy.
	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	Preferably we have a date. The minimum is a year and if applicable a season. The chance that we are outside the growing season for crop type (CT) would be limited. However, in case of a large country and limited info on crop calendars there could be a too large bias introduced especially in case of multiple seasons. For land cover (LC) this is less critical, at least we need a year.
	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 but most likely only one season[1]	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[1] e.g. first season always wheat and second season always rice	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[1] e.g. both seasons have rice or seasons can have rice or another crop but the order can change from year to year	Reject		
	Case 5 for LC: In case of land cover (LC) the absence of season info is not a problem	100		
Validation applied?	Yes[2]	100	25	Assume that most observations

[1] The issue of missing season information is that there is the risk of more seasons of the same crop and then the wrong validity time (centre date of the wrong crop calendar) might be used.
[2] The data set might have issues which could "degrade" the data set.

	No (doubtful)	80		are correct even if there was no validation, so weight is relatively small

Details calculations for Classification or Validation by crowd or expert (CV) e.g. Geo-Wiki and LACO-Wiki (at dataset level)

Quality Category	Description	Score range	Weight (%)	Justification
Geometry	Point/polygon (m)[3] (derived by expert)	100	40	We assume that the vector point/polygon are produced with high accuracy. For the raster datasets, the coarser the resolution the higher the chance of mixed pixels.
	Point/polygon (m)[3] (drawn by user)	80		
	Grid/Pixel 0-10 m	100		
	Grid/Pixel 11-20 m	80		
	Grid/Pixel 21-30 m	50		
	Grid/Pixel 31-50 m	20		
	Grid/Pixel > 50 m	Reject		
Level of accuracy of time	Imagery time	100	25	There is the risk that the submission date deviates too much from the imagery date with the consequence that the date is outside the targeted season
	Derived from submission date	50		
Average User Confidence[4]	>90	100	35	In general, we believe the visual interpretation is done thoroughly so we should not decrease the total score too much
	80-90	80		
	70-80	70		
	60-70	60		
	<=60	50		

[3] Point/polygon (m) is obtained from satellite image digitization. We assume that geometry is accurate.

[4] User confidence score is the data set average, calculated from the per-sample values based on method developed by IIASA. See Annex I.

Details calculations for Automated Classification (classified map) (at dataset level)

Quality Category	Description	Score range	Weight (%)	Justification
Geometry	Grid/Pixel 0-10 m	100	35	The coarser the resolution the higher the chance of mixed pixels.
	Grid/Pixel 11-20 m	80		
	Grid/Pixel 21-30 m	50		
	Grid/Pixel 31-50 m	20		
	Grid/Pixel > 50 m	Reject		
Level of accuracy of time	Real date	100	25	Preferably we have a date. The minimum is a year and if applicable a season. The chance that we are outside the growing season for crop type (CT) would be limited. However, in case of a large country and limited info on crop calendars there could be a too large bias introduced especially in case of multiple seasons. For land cover (LC) this is less critical, at least we need a year.
	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 but most likely only one season	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[1] e.g. first season always wheat and second season always rice	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[1] e.g. both seasons have rice or seasons can have rice or another crop but the order can change from year to year	Reject		
	Case 5 for LC: In case of land cover (LC) the absence of season info is not a problem	100		
Classification accuracy	More than 90%	100	40	Relatively important because the classification is the "pseudo observation". Less than 50% classification accuracy doesn't fit for our purpose.
	Between 80-90%	90		
	Between 70-80%	80		
	Between 60-70%	50		
	Between 50-60%	20		
	Less than 50%	Reject		

Details calculations for Formal Declaration (parcel registrations) (at dataset level)

Quality Category	Description	Score range	Weight (%)	Justification
Geometry	Polygon (m)[3]	100	40	We assume that the parcel registration information comes from the government and is accurate.
Level of accuracy of time	Real date	100	30	Preferably we have a date. The minimum is a year and if applicable a season. The chance that we are outside the growing season for crop type (CT) would be limited. However, in case of a large country and limited info on crop calendars there could be a too large bias introduced especially in case of multiple seasons. For land cover (LC) this is less critical, at least we need a year.
	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 but most likely only one season	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[1] e.g. first season always wheat and second season always rice	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[1] e.g. both seasons have rice or seasons can have rice or another crop but the order can change from year to year	Reject		
	Case 5 for LC: In case of land cover (LC) the absence of season info is not a problem	100		
Thematic accuracy	Correct definition of declared crop	100	30	We assume that the parcel registration information comes from the government and is accurate.

Annex I Per-sample user confidence score for data type Classification or Validation by crowd or expert (CV) (Geo-Wiki, LACO-Wiki)

Usually in LACO-Wiki there is 1 validation per location. If there is more than one campaign or validation session for a given sample points/polygons/pixels we will have more than one validation submission per location. In LACO-Wiki we found that most locations have 2 submissions, and more than 2 submissions is less than 10%.

Based on our empirical evidence from past campaigns, we can derive the following rules for user confidence:

Step 1: if the user is considered:

a) Expert, then base confidence= 80%
b) Non-expert, then base confidence=50%

Step 2: if the total number of validations is:

a) 1, then final confidence depends on **step 1**
b) 2 or more (crowd or experts) then:
    - If 2 or more people disagree, final confidence is low (<50%)
    - If 1 person disagrees, and:
        - less than 6 people agree, then final confidence is low (<50%)
        - 6 or more people agree then:
            - If 6-7 people agree, then final confidence = 60%
            - If 8-9 people agree, then final confidence = 70%
            - If 10-13 people agree, then final confidence = 80%
            - If 14 or more people agree, then final confidence = 90%
    - If no one disagrees, then:
        - If validations are done only by experts, then:
            - If 2 people agree then confidence =90%
            - If 3 or more people agree then confidence =95%
        - If validations are done only by non-experts, then:
            - If 2-3 people agree then confidence= 70%
            - If 4-5 people agree then confidence= 80%
            - If 6 or more people agree then confidence is 90%
        - If validations are done by a mix of experts and non-experts then:
            - If one expert and >=1 non-expert agree then confidence = 90%
            - If 2 or more experts agree, irrespective of number of non-experts confidence=90%

Expert/Non-expert considerations:

- Data from Geo-Wiki is usually considered to come from the general crowd, i.e. non-experts
- Data from LACO-Wiki can be considered coming from experts if the usernames/emails can be recognized as experts or the campaign can be recognized as an expert-led campaign, examples: Corine, local components, EU programs, specific username/email. Otherwise users are considered as non-experts.