



# WorldCereal MOOC I: Reference data for crop type mapping



# Exercises 'Quality Assessment of Reference Data'

By Arun Pratihast & Hendrik Boogaard



# Background

WorldCereal has developed a generic scheme to assess the quality of the reference data by calculating confidence scores.

The scheme differentiates 4 different data types:

- Field Observation Survey (field visit)
- Virtual (and/or automated) Interpretation (by photo, high resolution imagery etc)
- Automated Classification (high-quality classified map)
- Formal Declaration (parcel registrations systems)

The schema includes the spatial, temporal, and thematic accuracy since these are essential aspects of crop mapping.

For more detailed information regarding the involved steps, and a copy of the sheet which is used to compute these dataset quality scores, we refer to the following documents in the supporting data of this exercise:

"WorldCereal\_ConfidenceScoreCalculations\_v1\_1.pdf"

#### "WorldCereal\_DataConfidenceScore\_Calculator\_v3\_0.xlsx"

This quality assessment is typically performed **by the WorldCereal moderator** when users decide to make their data publicly available through the <u>WorldCereal Reference Data Module</u> (RDM). Each publicly available dataset in the RDM is characterized by a quality score related to land cover, crop type and/or irrigation information. The following exercises are meant to show how WorldCereal moderators manage the quality of reference data and which factors are typically involved in this process.

Please note that a user of the WorldCereal system does not need to go through these steps before being able to upload and work with their data in the system. These exercises are mainly meant to raise awareness on the aspect of data quality. As such, going through these exercises could make you decide to first work on your dataset to improve its quality before uploading the data into the WorldCereal RDM.

### Objectives

In this series of practical exercises you will learn:

- To investigate spatial, temporal, and thematic accuracy of a dataset
- To estimate the final confidence score of a dataset

The analysis will be conducted within the QGIS environment.

### **Exercise 1: Investigating Spatial Accuracy**

WorldCereal uses seven steps to access the spatial accuracy of point data as shown in the following figure:



Step 1: Obtain in-situ reference data

For the exercises, you will use a subset of the Tanzania Soil Information Service (TanSIS) dataset, as published by Walsh et al. (2017)<sup>1</sup>. The primary objective of this dataset is to develop and codify information on soils, cropping and pastoral systems, and landscapes to evaluate cropland productivity and soil fertility in Tanzania.

The data is available as a Geopackage file named '2017\_tza\_afsis\_point\_110'. The observed locations are presented in the following figure:



Step 2: Select the cropland observations

Open the attribute table and examine the field that contains the crop names. It has 230 data points. The 'Landcover' field includes crops names and 'No information'.

<sup>&</sup>lt;sup>1</sup> Walsh, Markus, Joel Meliyo, Bruce Scott, Barbara Walsh, and Bob Macmillan. 2021. "Tanzania Soil Information Service (TanSIS)." OSF. September 6. doi:10.17605/OSF.IO/4NGAU.

Q	2017_tza_afsis_poir	nt_110 — Feature	es Total: 230, Filter	red: 230, Select	ed: 0																							
/			8 8 5	🔍 🕆 📼	🏘 🗭 🛯 📾	1 🗮 🖷	Q, 6	7																				
	fid	fid_1	today	surveyor	lat	lon	maize	barley	milet	wheat	sorghum	rice	tef	ceother	r bean	chickpes	a cowpea	pidgeonpea	soybean	groundnut	leother	ROP	potato	spotato	cassava	roother	OCP	Landcover *
1	56	1017	3 8/4/2017	OMAR	-7.970666484	31.70379507	N	N	N	Y	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	Wheat
2	132	1024	9 15/06/17	Sv	-7.704791832	35.49854059	N	N	N	Y	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	Wheat
3	81	1019	8 13/04/17	Essau	-8.578041833	32.83687153	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	Y	N	Y	N	N	Y	Spotato
4	102	1021	9 19/04/17	Mk	-8.816160464	34.63065487	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	Y	N	Y	N	N	N	Spotato
5	166	1028	3 17/06/17	Ben & kiriba	-5.112324673	34.67652381	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	Y	N	Y	N	N	Y	Spotato
6	33	1015	0 1/1/2017	Vidky	-4.16620681000	34.68135047	N	N	N	N	Y	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	Y	Sorghum
7	34	1015	1 1/1/2017	Vicky	-4.16791029	34.67999553	N	N	N	N	Y	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	Y	Sorghum
8	96	1021	3 16/04/17	Essau	-8.735560931	34.20670904	N	N	N	N	Y	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	Sorghum
9	97	1021	4 16/04/17	Essau	-8.75838695600	34.24904461	N	N	N	N	Y	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	Y	Sorghum
10	144	1026	1 15/06/17	Mk	-6.877120373	36.13132149	N	N	N	N	Y	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	Sorghum
11	159	1027	6 17/06/17	Essau	-6.532801897	37.24632799	N	N	N	N	N	Y	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	Rice
12	36	1015	3 4/4/2017	Essau	-8.288403681	31.56159311	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	Y	No information
13	42	1015	9 5/4/2017	Grp 2	-7.168681731	31.01571611	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	No information
14	47	1016	4 7/4/2017	Essau	-8.306594789	31.28913748	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	No information
15	52	1016	9 7/4/2017	OMAR	-8.238301601	31.82355773	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	No information
16	57	1017	4 8/4/2017	Grp 2	-7.68572684	31.16029427	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	Y	No information

Select all cropland using the Query Builder. The following query can be used to select cropland data:

eids				Values
millet				Q Search
wheat				
sorghum				Bean
rice				Leother
tef				Maize
ceother				Millet
bean				Mixed Croppin
chickpea				No information
cowpea				Rice
pidgeonpea				Sorghum
soybean				Spotato
groundnut				Wheat
leother				
ROP				
potato				
spotato				
roother				
OCP				
Landcover			-	Lice unfilte
Landcover			•	Use unlitte
Operators				
-	<	>	LIKE	%
<=	>=	!=	ILIKE	AND

In this dataset, there are 21 data points that have no crop information and 209 points with crop information.

Save the cropland data in the Geopackage file '2017\_tza\_afsis\_point\_110\_crop.gpkg'.

#### **Step 3: Download reference layers**

You will use OpenStreetMap (OSM) datasets for this purpose. OSM data can be downloaded from Geofabrik: (<u>https://download.geofabrik.de/</u>). QGIS also offers the QuickOSM plugin (<u>https://plugins.qgis.org/plugins/QuickOSM/</u>), which allows you to download data directly in QGIS.

For this exercise, you can use the OSM road layers for Tanzania already prepared for the MOOC: tza\_osm\_road.gpkg.

#### Step 4: Perform spatial join analysis

The aim of this step is to determine whether a point is located properly within the field or if the point is too close to, or on, a road. To achieve this, you will use the spatial join operation in QGIS. For analysis within a single country, use the country's local or national projection system to minimize distortion. For larger regions like Africa, avoid UTM zones due to inconsistencies when crossing zones and instead use a continent-wide projection, such as Africa Albers Equal Area Conic, for consistent and accurate measurements. Ensure both the point dataset and the road layer are in the same projection before performing the spatial join to avoid mismatches and ensure accurate results.

The datasets are about Tanzania, hence we will use UTM zone 36S.

Use "Assign projection" tool.



Save the data as Geopackage files '2017\_tza\_afsis\_point\_110\_crop\_UTM' and 'tza\_osm\_road\_UTM'.

Then, use the '*Join attributes by nearest* ' tool and save the data as Geopackage file '2017\_tza\_afsis\_point\_110\_Crop\_UTM\_join'.

🞗 Vector General - Join Attributes by Nearest
Parameters Log
Input layer
2017_tza_afsis_point_110_Crop_UTM.gpkg [EPSG:32736]
Selected features only
Input layer 2
√° tza_osm_road_UTM.gpkg [EPSG:32736]
Selected features only
Layer 2 fields to copy (leave empty to copy all fields) [optional]
0 fields selected
Discard records which could not be joined
Joined field prefix [optional]
distance
Maximum nearest neighbors
1
Maximum distance [optional]
Not set
Joined layer [optional]
D:/WorldCereal_phase2/MOOC_ex4/2017_tza_afsis_point_110_Crop_UTM_join.gpkg
✓ Open output file after running algorithm
Unjoinable features from first layer [optional]
[Skip output]
Open output file after running algorithm

#### Step 5: Filter points near road

Next, you will select the points located within crop fields. For this purpose, points near roads will be excluded. Specifically, points within a 20-meter distance from the road will be rejected, while points further than 20 meters will be selected.

ds			Values			
123 distancelayer		<b></b>	Q. Search			
abc distancebridge			0.6233276515	507199		
			2.4749425683	583186		
abc distancetunnel			3.3010432053	57813		
123 n			4.0832267539	87609		
			4.1998795857	73884		
1.2 distance			4.6501339935	66814		
1.2 feature_x			5.8055672890	25199		_
12 ( )			7.3000042111	82379		
1.2 feature_y			7.6568626979	27328		
1.2 nearest_x			8.9630220581	31223		-
1.2 nearest v			Sa	mple	All	
/						
		•	Use unfilter	ed layer		
Operators						
= <	>	LIKE	%	IN	NOT IN	
<= >=	!=	ILIKE	AND	OR	NOT	
vider Specific Filter Expressio	'n					



In this dataset, 17 data points are close to road. Figure shows the rejected data points:

Further investigation of the data point using high-resolution Google Satellite data clearly indicates that the data point is located on the road.



Data points located more than 20 meters from the road are considered reliable cropland observations. In this case, you will select 192 data points and save these selected data points as Geopackage file '2017\_tza\_afsis\_point\_crop\_UTM\_accepted.

#### Step 6: Random selection of points for visual interpretation.

To gain an understanding of the data quality of the remaining data, we will randomly select 5% of the data for further visual inspection. Use the Random Extract tool in QGIS to randomly select 5% of the data points:

Processing Toolbox
- 🎇 👶 🛈 🖹   🖻   🔧 -
random     Recently used
Random extract
Random extract within subsets
<ul> <li>Random selection within subsets</li> <li>Q Raster creation</li> </ul>
Create random raster layer (binomial distribution)
Create random raster layer (exponential distribution) Create random raster layer (gamma distribution)
Veate random raster layer (geometric distribution)
Create random raster layer (negative binomial distribution)
Create random raster layer (normal distribution)
Create random raster layer (uniform distribution)
Q Vector creation
Random points in extent
Random points in layer bounds
Random points in polygons     Random points inside polygons
)) Random points on lines
Q Vector selection
* Random extract within subsets
-7 Random selection
Provide the section within subsets     Section within subsets     Section within subsets     Section within subsets
▼ Raster (r.*)
₩ r.random ŵ r.random.cells
💮 r.random.surface
₩ v.kcv
v.perturb
X V.Iandom
🔇 Vector Selection - Random Extract
Parameters Log
Input layer
2017 tza afsis point vi [EPSG:32736]
[ ·
Selected features only
Method
Piction
Percentage of features
Number/percentage of features
5
S
2017_tza_afsis_point_vi
✓ Open output file after running algorithm

Save the sample as Geopackage file '2017\_tza\_afsis\_point\_vi'.

Export this sample as KML format.

Save Vecto	r Layer as				2				
Format	Keyhole Markup La	inguage [KML]			•				
File name	2017_tza_afsis_point_vi								
Layer name 2017_tza_afsis_point_vi									
CRS	EPSG:32736 - WG	S 84 / UTM zone 36S			•				
Encoding		UTF-8							
Save only	selected features								
Select fi	elds to export a	nd their export options							
✔ Persist lay	er metadata								
Symbology ex	port	No Syn	nbology		•				
Scale		1:1000	000		- 🔊 -				
▼ Geomet	ry								
Geometry t	ype	Autom	atic		-				
Force m	ulti-type								
Include	z-dimension								
Exte	nt (current: non	e)							
AltitudeMod	le relativeToG	round			•				
DOCUMENT	_ID root_doc								
Description	Field Description								
NameField	Name								
		✓ Add saved file to map	ОК	Cancel	Help				

Detailed guidelines involving visual inspection of samples can be found in the following document, contained within the supporting materials of this exercise:

"Guidelines\_Visual\_Inspection\_Samples\_v1\_0.pdf"

Load the Dataset into Google Earth Engine and perform visual interpretation of the sample dataset. For example, the location below is perfect in the crop field.



Add in description "Yes" if mapping is accurate Else "No".

2				
ame:				
Description	Style, Color	View	Altitude	
Add lin	k Add w	eb image	Add local image	
Yes				

Visual inspection is subjective and often depends on the interpreter. In this case, nine points were evaluated, two of which were not in the crop field (see the example in the following figure).



Save your results and count the number of suspicious cases.

#### Step 7: Finally assign the spatial accuracy case

WorldCereal proposes different case scenarios (as shown in the table below) to value the analysis of visual interpretation.

Case	Description
Case 0	Expert evaluated samples of cleaned data show no issues
Case 1	Expert evaluated samples of cleaned data show issues (between 1-10%)
Case 2	Expert evaluated samples of cleaned data show issues (between 10-25%)
Case 3	Expert evaluated samples of cleaned data show issues (between 25-50%)
Case 4	Expert evaluated samples of cleaned data show issues (between 25-50%)

In the exercise above, there were 9 samples, 2 of which were found to be suspicious, placing this scenario under Case 2 (issues between 10-25%).

# Exercise 2: Investigating Temporal Accuracy

Open the attribute table of Geopackage '2017\_tza\_afsis\_point\_crop\_UTM\_accepted'.

Examine attributes to find the date when the crop was observed.

Q	<b>Q</b> 2017_tza_afsis_point_110 — Features Total: 230, Filtered: 230, Selected: 0 — X													
/		1 <b>6</b> × 6 6	1 🗧 🗮 💟	💊 🍸 🔳	🏘 🗭 i 🖪 🛯	1 📰 📰	Q. 🗇							
_	fid	fid_1	today	surveyor	lat	lon	maize	barley	millet	wheat	sorghum	rice		•
1	56	10173	8/4/2017	OMAR	-7.970666484	31.70379507	N	N	N	Y	N	Ν	N	
2	132	10249	15/06/17	Sv	-7.704791832	35.49854059	N	N	N	Y	N	Ν	N	
3	81	10198	13/04/17	Essau	-8.578041833	32.83687153	N	N	N	N	N	N	N	
4	102	10219	19/04/17	Mk	-8.816160464	34.63065487	N	N	N	N	Ν	Ν	N	
5	166	10283	17/06/17	Ben & kiriba	-5.112324673	34.67652381	N	N	N	N	N	N	N	
6	33	10150	1/1/2017	Vicky	-4.16620681000	34.68135047	N	N	N	N	Y	Ν	N	
7	34	10151	1/1/2017	Vicky	-4.16791029	34.67999553	N	N	N	N	Y	N	N	
8	96	10213	16/04/17	Essau	-8.735560931	34.20670904	N	N	N	N	Y	Ν	N	
9	97	10214	16/04/17	Essau	-8.75838695600	34.24904461	N	N	N	N	Y	N	N	
10	144	10261	15/06/17	Mk	-6.877120373	36.13132149	N	N	N	N	Y	Ν	N	
11	159	10276	17/06/17	Essau	-6.532801897	37.24632799	N	N	N	N	N	Y	N	
12	36	10153	4/4/2017	Essau	-8.288403681	31.56159311	N	N	N	N	N	Ν	N	
13	42	10159	5/4/2017	Grp 2	-7.168681731	31.01571611	N	N	N	N	N	N	N	
14	47	10164	7/4/2017	Essau	-8.306594789	31.28913748	N	N	N	N	N	Ν	N	
15	52	10169	7/4/2017	OMAR	-8.238301601	31.82355773	N	N	N	N	N	N	N	
16	57	10174	8/4/2017	Grp 2	-7.68572684	31.16029427	N	N	N	N	Ν	Ν	N	Ŧ
4													•	
<b>1</b> S	how All Features 🖕												3	1

# Exercise 3: Overall dataset confidence score calculation

Open the confidence calculator Excel file, supplied in the supporting material to this exercise:

"WorldCereal\_DataConfidenceScore\_Calculator\_v3\_0.xlsx"

Evaluate the dataset:

- Spatial Accuracy:
  - No GPS information
  - Spatial context analysis: Case 2 (Expert evaluated samples of cleaned data show issues (between 10-25%)).
- Temporal Accuracy: survey date is present
- Validation: was done by the original data holder (see 10.17605/OSF.IO/4NGAU with information on raw and tidy files and the ODK form Crop\_scout.xlsx for info on spatial accuracy)

Calculate the confidence score (e.g., 82.8%).

FieldObservationSurvey / windshiel	d (at dataset level)			
Quality Category	Description	Score & Reduction factor	Weight (%)	Total Score
Geometry (spatial accuracy based on	If GPS info is not present	95	40	22.8
GPS)				
Geometry (spatial context analysis	Case 2: Evaluated samples of cleaned data show issues (between	0.4		
by benchmarking against non-arable	10-25%)			
spatial features e.g., roads, water				
bodies, railway, buildings, nature				
areas etc.)				
Level of accuracy of time	Real date	100	35	35
Validation applied	Yes	100	25	25
	Grand Total Confidence Score			82.8

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

# Summary

In the above exercises, you learnt which factors are typically involved when WorldCereal moderators are evaluating the quality of a reference dataset and how to perform a quality assessment of point data. The steps for polygon data are similar, with the addition of more detailed analysis of polygon geometry, the heterogeneity of mapped polygons within cropland, and performing an intersection analysis with roads.