WorldCereal MOOC I:
Reference data for crop type mapping

Exercises 'Reference data harmonization and cleaning'

By Hendrik Boogaard & Arun Pratihast

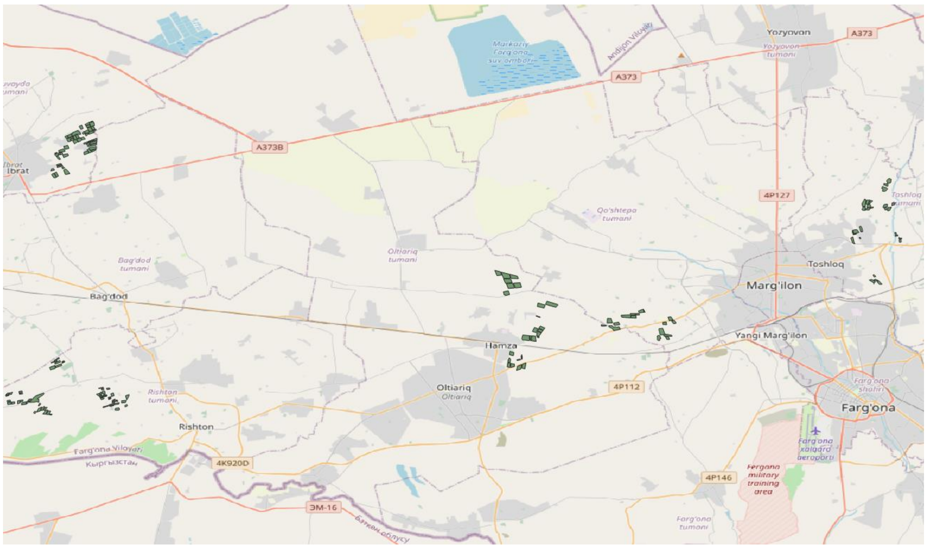
Background

When working with reference data in the WorldCereal processing system the data should comply to the following specifications:

- Data must contain spatial geometry (points or polygons)
- Data must include land cover and/or crop type observations[1]
- Data must have the date when the declared land cover or crop type was present at the specified location[2]
- A validity time of observations later than January 1st 2017[3]

In case you have a simple point or polygon dataset with only one attribute describing land cover and/or crop type and one attribute with the observation date, you can go directly to Chapter 5 (Introduction to the WorldCereal Reference Data Module (RDM)) and learn how you can easily harmonize and upload your data into the WorldCereal system (https://ewoc-rdm-ui.iiasa.ac.at).

However, your dataset can be more complicated. In such a case we suggest reviewing your dataset and checking if some data preparation might be required before uploading your data into the WorldCereal RDM. The following exercises are focused on this data preparation to get the most out of your data and avoid possible mistakes. For the exercises we use a subset of the data published by Remelgado et al. (2020)[4]. The original data covers multiple years (2008-2018) and a wide range of cropping systems. We limited the dataset to harvest year 2018, the Fergana region and a limited selection of crops and cropping systems. The observed locations are presented in the following figure:



The data for the exercises are available as a Geopackage file named 'cawa_2018_crop_selection.gpkg'.

[1] Land cover and crop type information should be combined into a single dataset attribute with data type string

[2] It can be specified in one of two ways: (1) separately for each individual observation (in which case it should be included as a dataset attribute in string or date format), or (2) as a single date for the entire dataset (which can be selected during the upload into the WorldCereal Reference Data Module.

[3] Due to reduced availability of satellite data prior to 2017, we currently do not support datasets before 2017

[4] 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. https://doi.org/10.1038/s41597-020-00591-2 (data via figshare: https://doi.org/10.6084/m9.figshare.12047478.v2)

Objectives

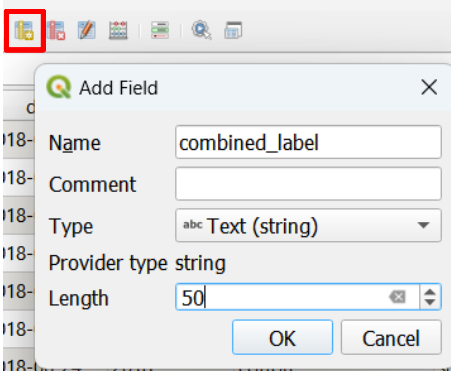
In this series of practical exercises you will learn:

- To refine the mapping to the WorldCereal land cover/crop type legend
- To deal with specific situations
- To estimate a date if missing

The analysis will be conducted within the QGIS environment.

Exercise 1: Refine mapping

- Open file 'cawa_2018_crop_selection.gpkg' in Q-GIS
- Next, open the attribute table
- Identify the attribute holding land cover and/or crop type observations
- One of the crop type labels can be refined with the help of another attribute. Which attribute and crop type would that be?
- In the upload procedure of the WorldCereal Reference Data Module (RDM) you must indicate the attribute holding the land cover and/or crop type observations (label_1). Next, the upload procedure does an automated AI-assisted mapping to the WorldCereal legend[5]. To profit from the information provided by attribute label_2 you would need to merge the information of the attributes label_1 and label_2. Therefore, you need to do the following:
    - Enable editing the attribute table
    - Add a new attribute, 'combined_label' (datatype string, length 50):



Add Field

Name: combined_label

Comment:

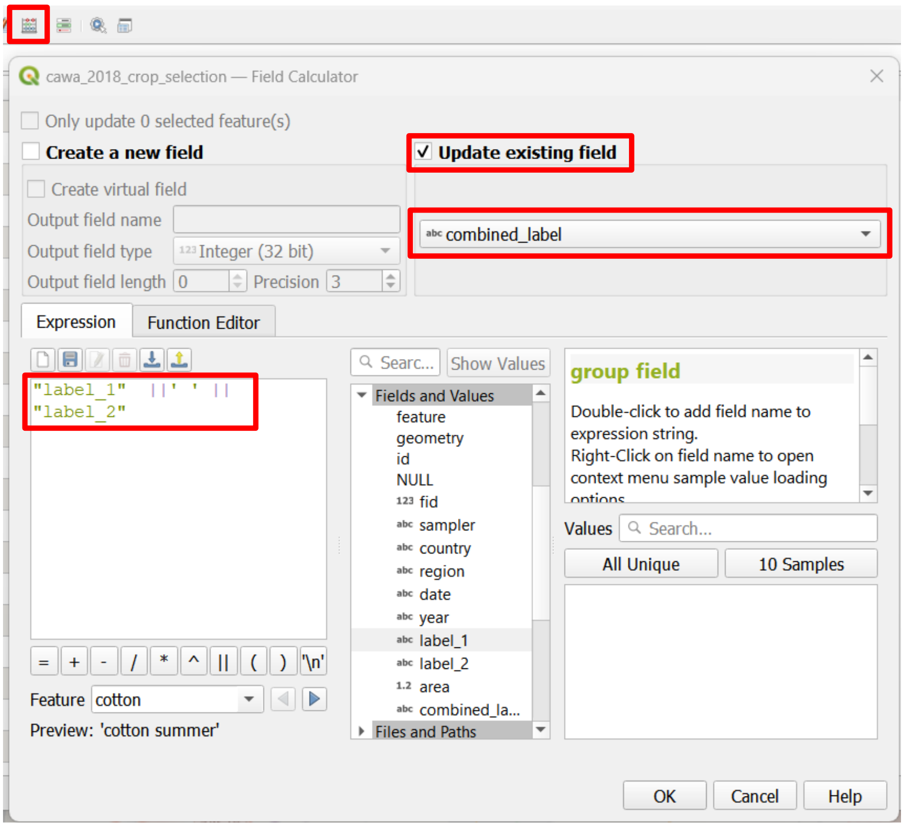
Type: abc Text (string)

Provider type: string

Length: 50

    - Open the field calculator, enable 'Update existing field', select attribute 'combined_label' and define the expression: "label_1" ||' ' || "label_2"

[5] The WorldCereal legend can be found via the following URL: https://artifactory.vgt.vito.be/artifactory/auxdata-public/worldcereal/legend/WorldCereal_LC_CT_legend_latest.pdf



cawa_2018_crop_selection — Field Calculator

□ Only update 0 selected feature(s)

□ Create a new field

☑ Update existing field

□ Create virtual field

Output field name

Output field type: Integer (32 bit)

Output field length: 0 Precision: 3

combined_label

```
"label_1"  ||' ' ||
"label_2"
```

Feature: cotton

Preview: 'cotton summer'

group field

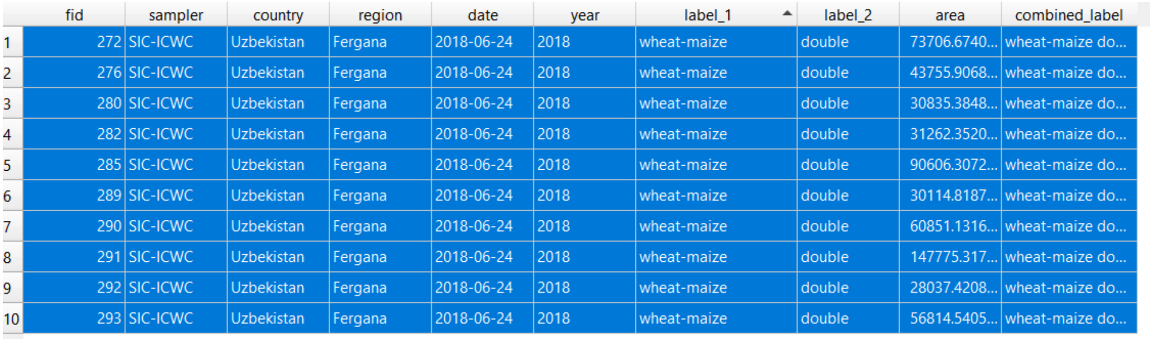
Double-click to add field name to expression string.
Right-Click on field name to open context menu sample value loading options

- ○ Save the edits!

- Now there is the new attribute ‘combined_label’ holding information that the AI-assisted mapping in the upload procedure of the RDM can use to propose the most detailed mapping for WorldCereal. Of course, you would still need to check the suggested mapping using the legend available via this link.

Exercise 2: Deal with specific situations

Before uploading a data set into the WorldCereal RDM you would need to be sure that all data are logical, consistent and as detailed as possible. Of course, it is not possible in this MOOC to discuss all situations that could emerge. As an example, we have observations in the file cawa_2018_crop_selection.gpkg that needs specific attention:

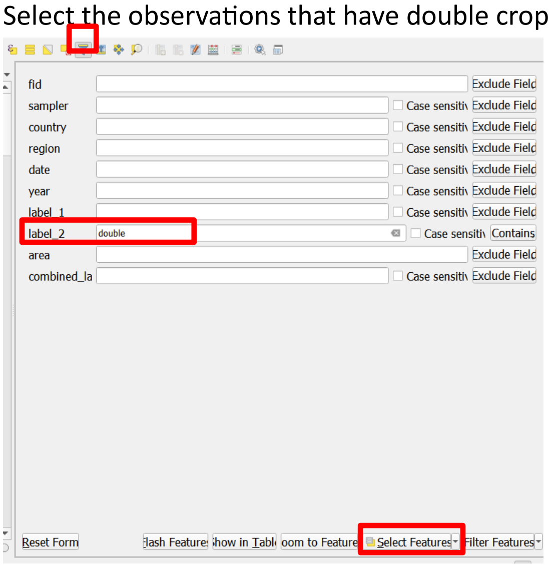


	fid	sampler	country	region	date	year	label_1	label_2	area	combined_label
1	272	SIC-ICWC	Uzbekistan	Fergana	2018-06-24	2018	wheat-maize	double	73706.6740...	wheat-maize do...
2	276	SIC-ICWC	Uzbekistan	Fergana	2018-06-24	2018	wheat-maize	double	43755.9068...	wheat-maize do...
3	280	SIC-ICWC	Uzbekistan	Fergana	2018-06-24	2018	wheat-maize	double	30835.3848...	wheat-maize do...
4	282	SIC-ICWC	Uzbekistan	Fergana	2018-06-24	2018	wheat-maize	double	31262.3520...	wheat-maize do...
5	285	SIC-ICWC	Uzbekistan	Fergana	2018-06-24	2018	wheat-maize	double	90606.3072...	wheat-maize do...
6	289	SIC-ICWC	Uzbekistan	Fergana	2018-06-24	2018	wheat-maize	double	30114.8187...	wheat-maize do...
7	290	SIC-ICWC	Uzbekistan	Fergana	2018-06-24	2018	wheat-maize	double	60851.1316...	wheat-maize do...
8	291	SIC-ICWC	Uzbekistan	Fergana	2018-06-24	2018	wheat-maize	double	147775.317...	wheat-maize do...
9	292	SIC-ICWC	Uzbekistan	Fergana	2018-06-24	2018	wheat-maize	double	28037.4208...	wheat-maize do...
10	293	SIC-ICWC	Uzbekistan	Fergana	2018-06-24	2018	wheat-maize	double	56814.5405...	wheat-maize do...

These 10 observations have double cropping defined, thus a rotation of wheat (winter) and maize (summer). See also the paper for more background on the cropping systems in this region. So, each record has potentially two observations: an observation of winter wheat and an observation of summer maize. Without any intervention these observations will be mapped to mixed arable cropping, in more detail WorldCereal code 11-14-01-000-0 (hierarchical order: level 1: temporary_crops, level 2: mixed_arable_crops, level 3: cereal_mixed_with)[5]. This is wrong because these observations describe a crop rotation and not inter-cropping.

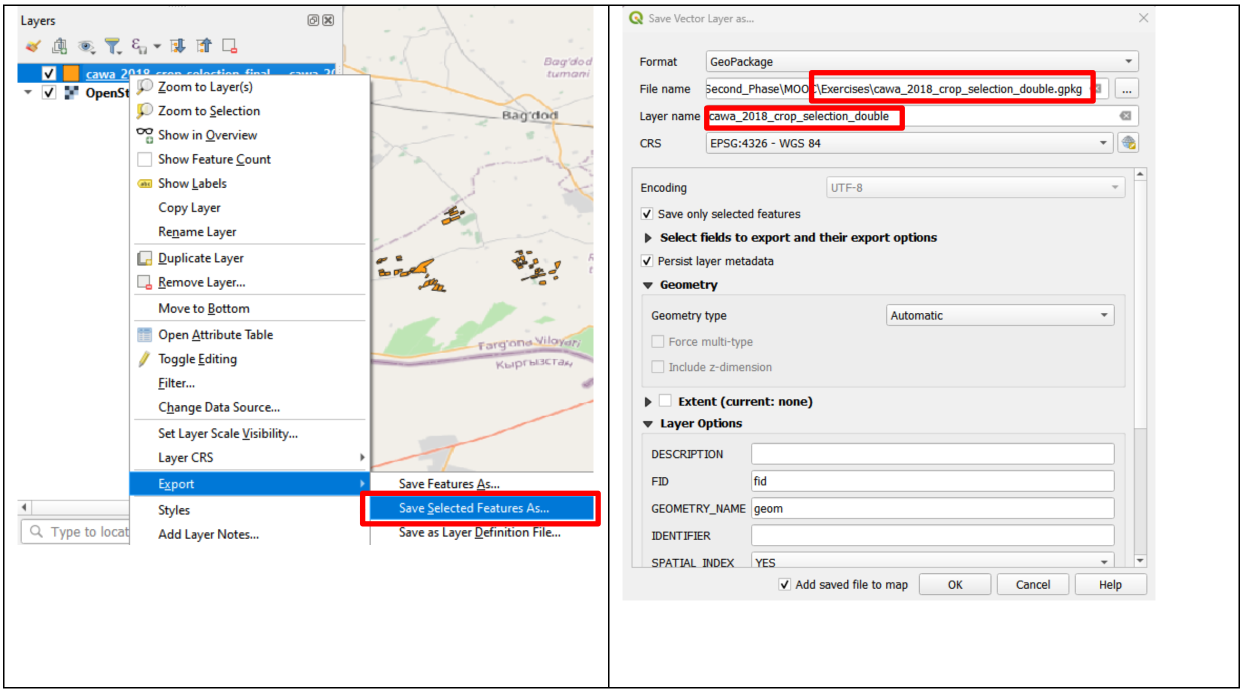
To fix this you would need to do copy the observations and assign winter wheat to one set of 10 records and assign maize to the other 10 records:

- Select the observations that have double cropping (10 observations)



fid
sampler
country
region
date
year
label_1
label_2 double ☐ Case sensiti Contains
area
combined_la

- Save selected records in separate Geopackage file 'cawa_2018_crop_selection_double' with a similar layer name



Format GeoPackage

File name Second_Phase\MOO...\Exercises\cawa_2018_crop_selection_double.gpkg

Layer name cawa_2018_crop_selection_double

CRS EPSG:4326 - WGS 84

Encoding UTF-8

☑ Save only selected features

▸ Select fields to export and their export options

☑ Persist layer metadata

▾ Geometry

Geometry type Automatic

□ Force multi-type

□ Include z-dimension

▸ □ Extent (current: none)

▾ Layer Options

DESCRIPTION

FID fid

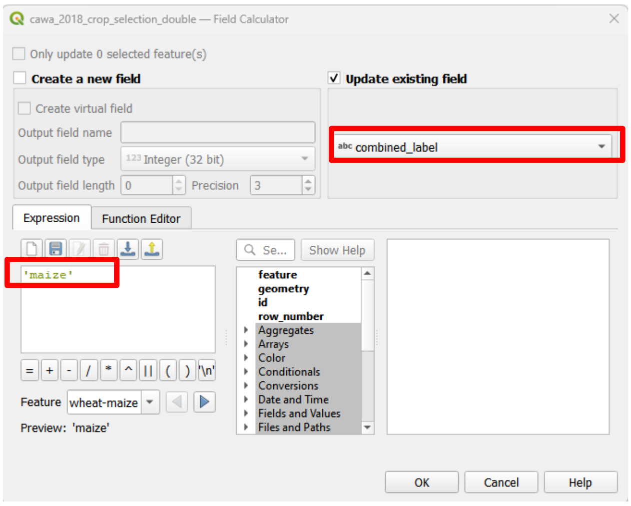
GEOMETRY_NAME geom

IDENTIFIER

SPATIAL_INDEX YES

☑ Add saved file to map

- Select layer ‘cawa_2018_crop_selection_double’ and open the attribute table
- Enable editing the attribute table
- Update attribute ‘combined_label’ to ‘maize’


cawa_2018_crop_selection_double — Field Calculator

□ Only update 0 selected feature(s)

□ Create a new field

☑ Update existing field

□ Create virtual field

Output field name

Output field type Integer (32 bit)

Output field length 0 Precision 3

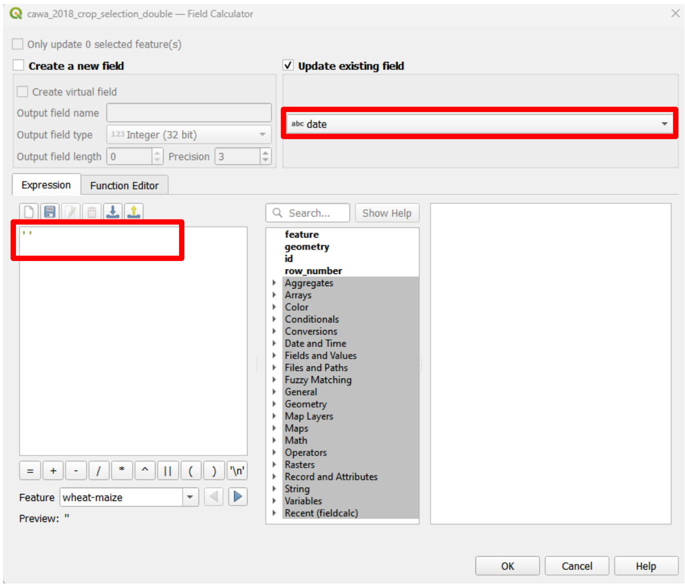
combined_label

Expression: 'maize'

Feature wheat-maize

Preview: 'maize'

- Update attribute ‘date’ and make it empty (note that in exercise 3 we will assign a logical date)



cawa_2018_crop_selection_double — Field Calculator

□ Only update 0 selected feature(s)

□ Create a new field

□ Create virtual field

Output field name

Output field type: Integer (32 bit)

Output field length: 0 Precision: 3

☑ Update existing field: date

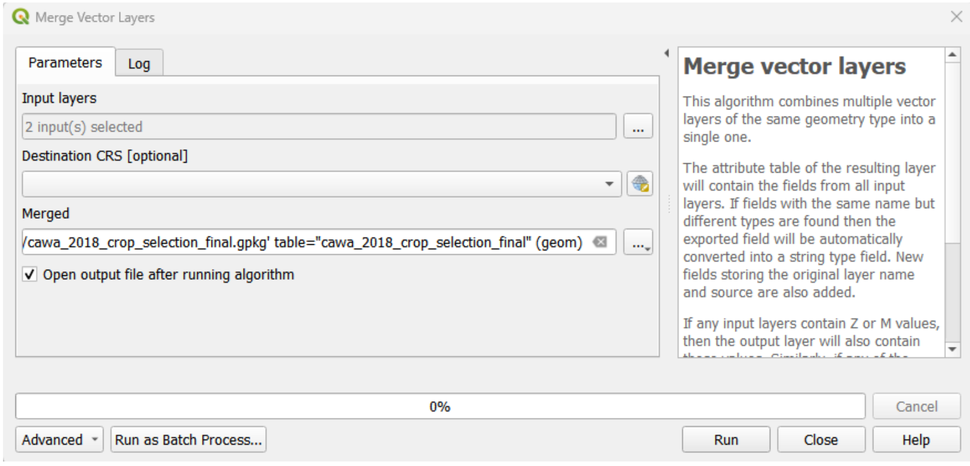
Expression:

''

Feature: wheat-maize

Preview: ''

- Save the edits!
- Select function ‘Merge vector layers’ in the Processing toolbox
- Select the layers ‘cawa_2018_crop_selection_double’ and ‘cawa_2018_crop_selection’ and save as Geopackage ‘cawa_2018_crop_selection_final’ with the same name for the layer



Merge Vector Layers

Input layers

2 input(s) selected

Destination CRS [optional]

Merged

/cawa_2018_crop_selection_final.gpkg' table="cawa_2018_crop_selection_final" (geom)

☑ Open output file after running algorithm

Merge vector layers

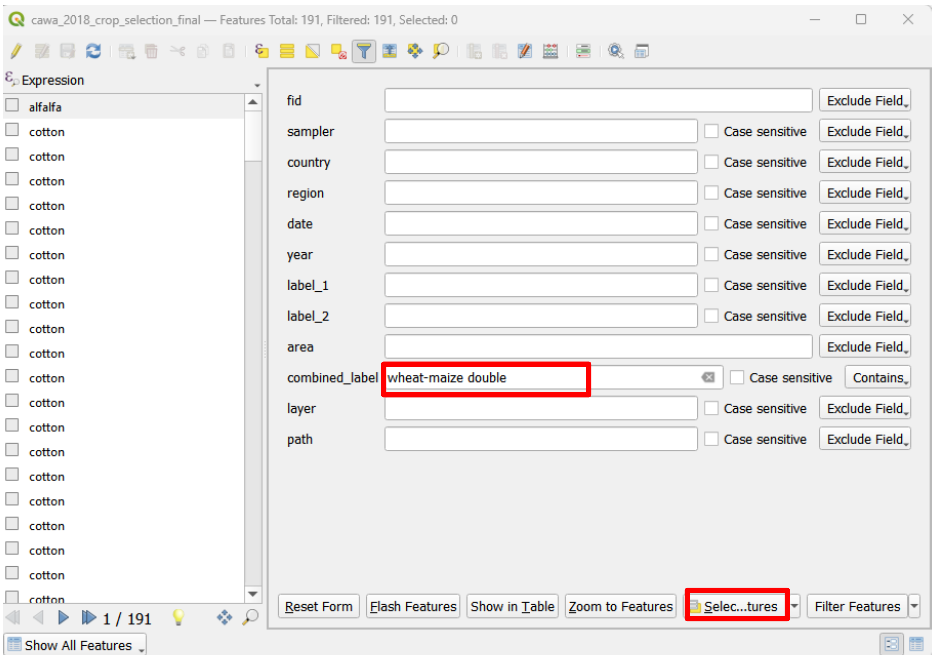
This algorithm combines multiple vector layers of the same geometry type into a single one.

The attribute table of the resulting layer will contain the fields from all input layers. If fields with the same name but different types are found then the exported field will be automatically converted into a string type field. New fields storing the original layer name and source are also added.

If any input layers contain Z or M values, then the output layer will also contain

0%

- Open the attribute table of layer ‘cawa_2018_crop_selection_final’
- Select the observations that have double cropping and that were not yet corrected (10 observations)

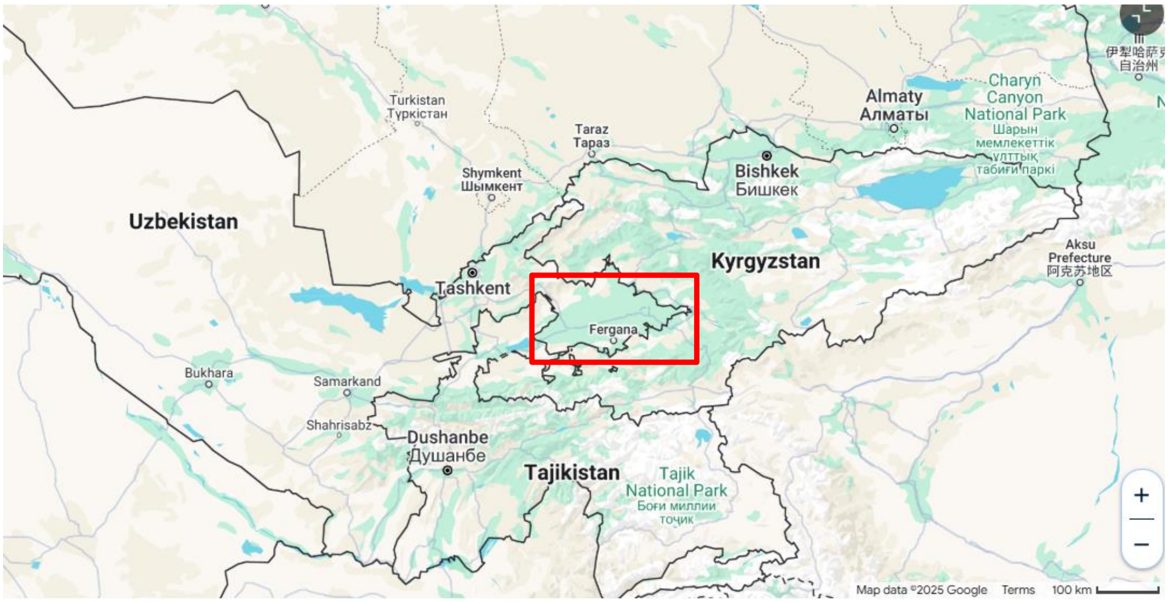

cawa_2018_crop_selection_final — Features Total: 191, Filtered: 191, Selected: 0

combined_label: wheat-maize double — □ Case sensitive — Contains

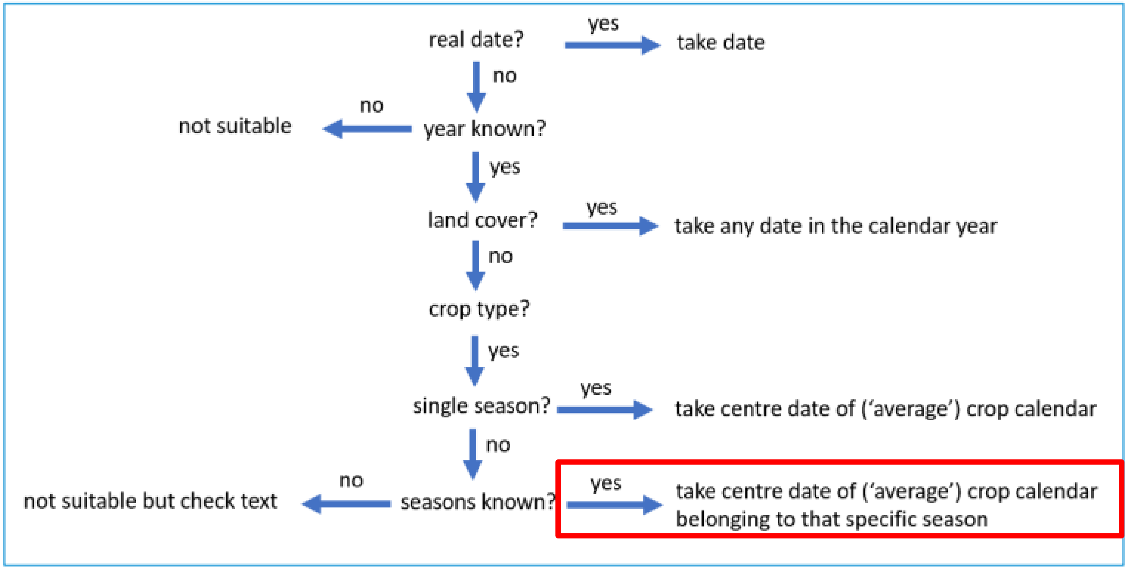
- Enable editing the attribute table
- Update attribute ‘combined_label’ to ‘winter wheat’
- Update attribute ‘date’ and make it empty (note that in exercise 3 we will assign a logical date)
- Save the edits!

Exercise 3: Estimate a date if missing

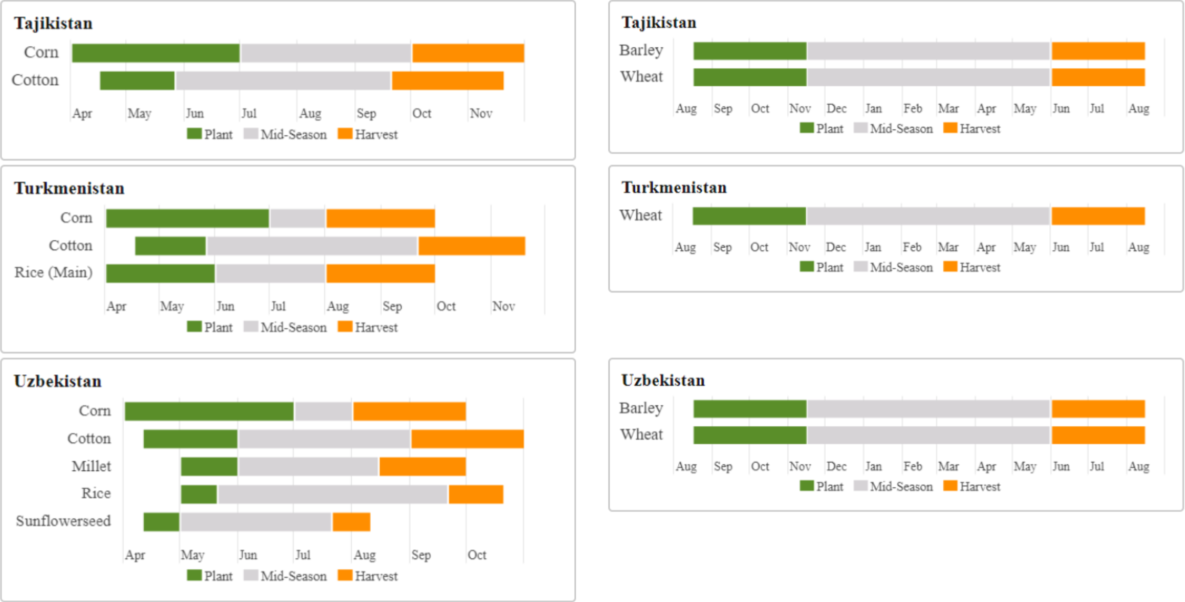
In exercise 2 you removed the dates of the observations with double cropping. However, you need a date for each observation. The date can be derived from the year and the governing crop calendar of the region of concern (see red square in the map below).



We have dedicated instructions that can be found here. In this case the following applies: 1) we are interested in crop types, 2) we know the year: 2018, and 3) there are two seasons: winter and summer. The scheme below illustrates the suggested strategy.

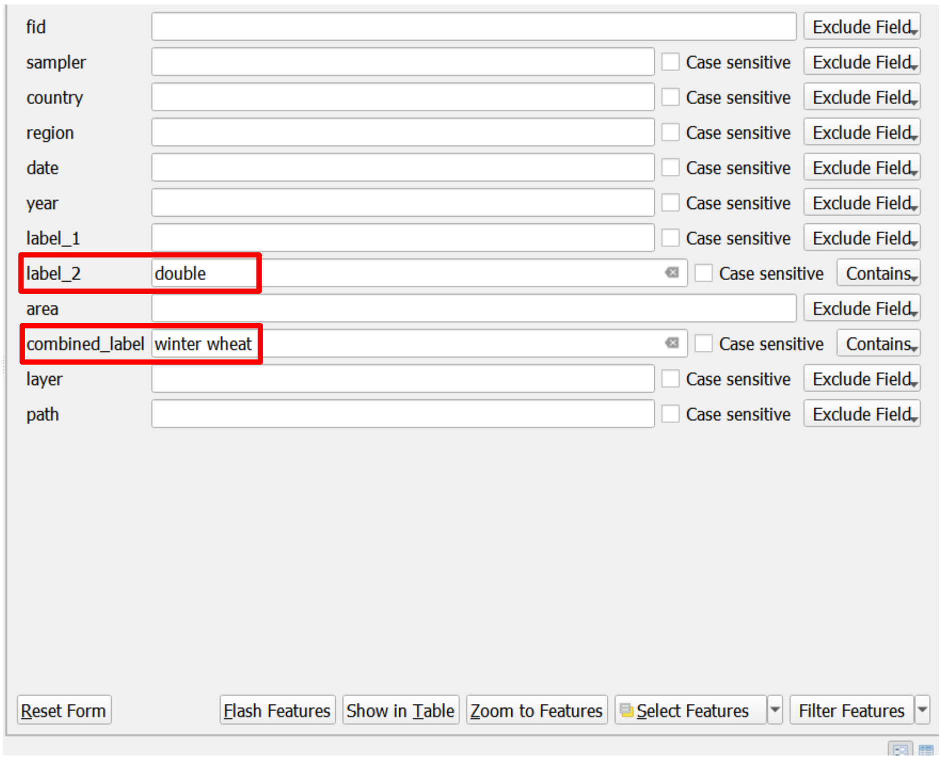

real date?
yes
take date
no
not suitable
no
year known?
yes
land cover?
yes
take any date in the calendar year
no
crop type?
yes
single season?
yes
take centre date of ('average') crop calendar
no
not suitable but check text
no
seasons known?
yes
take centre date of ('average') crop calendar belonging to that specific season

So, you must determine the center date of the crop calendar belonging to the specific season. You could consult the USDA crop calendars[6]:



To fill in the date for the 20 records, having a missing date, you would need to do the following:

- Select the observations that have missing dates by first selecting the records that have value ‘double’ for attribute ‘label_2’ and value ‘winter wheat’ for attribute ‘combined_label’



fid			Exclude Field
sampler		□ Case sensitive	Exclude Field
country		□ Case sensitive	Exclude Field
region		□ Case sensitive	Exclude Field
date		□ Case sensitive	Exclude Field
year		□ Case sensitive	Exclude Field
label_1		□ Case sensitive	Exclude Field
label_2	double	□ Case sensitive	Contains
area			Exclude Field
combined_label	winter wheat	□ Case sensitive	Contains
layer		□ Case sensitive	Exclude Field
path		□ Case sensitive	Exclude Field

[6] https://ipad.fas.usda.gov/rssiws/al/crop_calendar/stans.aspx

- Enable editing the attribute table
- Update attribute ‘date’ and set the value to your estimated center date of the winter wheat season (e.g. '2018-04-01')
- Next, select the remaining observations that have missing dates by selecting the records that have value ‘double’ for attribute ‘label_2’ and value ‘maize’ for attribute ‘combined_label’
- Update attribute ‘date’ and set the value to your estimated center date of the maize season (e.g. '2018-09-01')
- Save the edits!

Exercise 4: Check relevant metadata

By default, your dataset, uploaded in the WorldCereal RDM, is private and cannot be accessed by others. WorldCereal encourages you to share data preferably with the community or at least share data with the WorldCereal consortium. Once you share data, metadata is needed to properly understand, value, use and acknowledge the data. When sharing with the community the relevant metadata is listed here (see also the WorldCereal RDM User interface):

- URL of the formal entity (organization / project) behind the original dataset. Note that your name and e-mail is already known by the WorldCereal system upon uploading your own data.
- Reference to the original data, in case the original dataset was published e.g. in an open data journal or data repository.
- Usage license. Who can use the original data and what restrictions do exist? Suggested options:
  - Creative commons[7] C0 (No Rights Reserved);
  - Creative commons CC BY (Attribution);
  - Creative commons CC BY-SA (Attribution-ShareAlike);
  - Creative commons CC BY-NC (Attribution-NonCommercial);
  - Creative commons CC BY-NC-SA (Attribution-NonCommercial-ShareAlike);
  - Private (No redistribution, only for private use);
  - Other (to be defined by the user/owner).
- Reference to license or formal agreement concerning the original data. For example, https://creativecommons.org/about/cclicenses or a link/doc provided by the owner/user.
- Preferred citation when dataset is used by others.
- A brief description of the objective of the original dataset.
- Was a sampling design used to obtain the original data? Possible values: Yes, No, Unknown. If yes:
  - Information on the use and type of such sampling design.
- Was validation done on the original dataset? Possible values: Yes, No, Unknown. If yes:
  - In case of 'Virtual (and/or automated) Interpretation (by photo, HR imagery etc)' give more information on the validation e.g. interpretation of very-high resolution imagery and in-situ pictures; visual interpretation of dense time-series of satellite imagery; use of classified maps etc.
  - In case of 'Field Observation Survey (field visit)' give information on quality control, curation etc."
- In case the validity time has been derived, we would like to know how the date was determined:
  - Derived from year, season and crop calendar (year and season are avaible)
  - Derived from year, season and crop calendar (only year is available)
  - Derived from campaign validation time in combination with imagery time. Note this applies to 'Virtual (and/or automated) Interpretation (by photo, HR imagery etc)'

[7] https://creativecommons.org

- Derived from submission time by expert or crowd. Note this applies to ‘Virtual (and/or automated) Interpretation (by photo, HR imagery etc)’
- Unknown

Summary

In the above exercises, you learnt to critically review and prepare your reference data before uploading the reference data to the WorldCereal Reference Data Module. Specifically, the content of the attribute, used to map the observations to the WorldCereal legend, needs attention. It is important to check if the content might be mis-interpreted in the mapping and if the information of other attributes might be relevant for the mapping.

Answers Exercise 1 Refine mapping

- Attribute label_1 holds the land cover and/or crop type observations
- Attribute label_2 has additional information on seasonality and it can be used to refine the wheat label: winter wheat