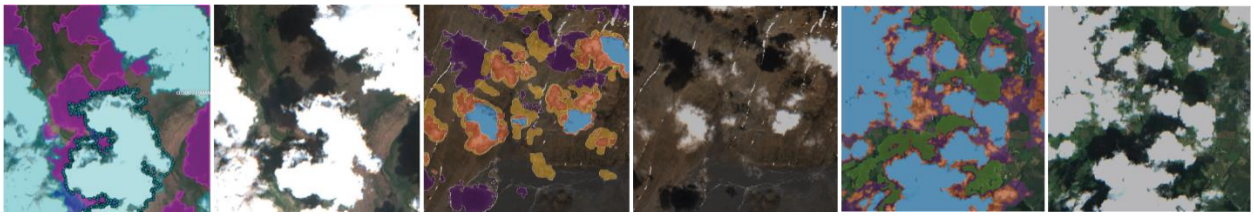
Sentinel-2 KappaZeta Cloud and Cloud Shadow Masks



1. General Information

The dataset consists of 4403 labelled subscenes from 155 Sentinel-2 (S2) Level-1C (L1C) products distributed over the Northern European terrestrial area. Each S2 product was oversampled at 10 m resolution for 512 x 512 pixels subscenes. 6 L1C S2 products were labelled fully. Among other 149 S2 products the most challenging ~10 subscenes per product were selected for labelling. In total the dataset represents 4403 labelled Sentinel-2 subscenes, where each sub-tile is 512 x 512 pixels at 10 m resolution. The dataset consists of around 30 S2 products per month from April to August and 3 S2 products per month for September and October. Each selected L1C S2 product represents different clouds, such as cumulus, stratus, or cirrus, which are spread over various geographical locations in Northern Europe (Figure 1).

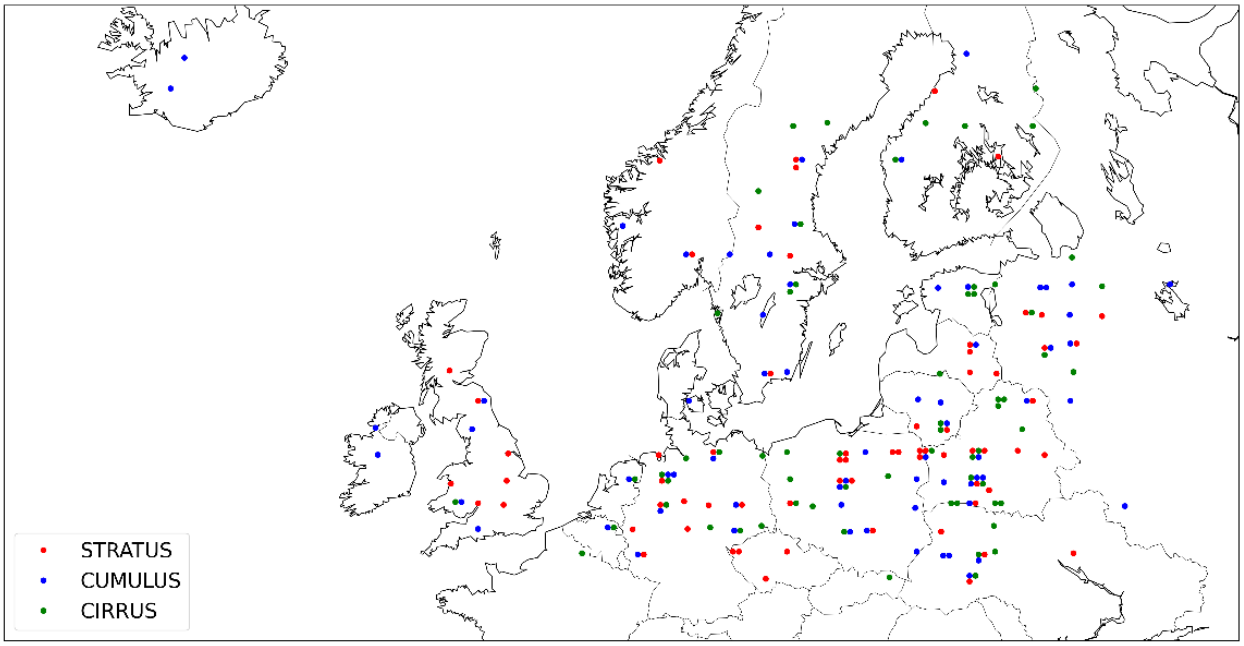


Figure 1. Sentinel-2 tiles used for labelling. Images with cumulus clouds are indicated as blue dots, images with stratus clouds are marked as red and images with cirrus clouds are marked as green. Each dot corresponds to one Sentinel-2 100x100 km data product.

The classification pixel-wise map consists of the following categories:

- 0 – MISSING: missing or invalid pixels;
- 1 – CLEAR: pixels without clouds or cloud shadows;
- 2 – CLOUD SHADOW: pixels with cloud shadows;
- 3 – SEMI TRANSPARENT CLOUD: pixels with thin clouds through which the land is visible; include cirrus clouds that are on the high cloud level (5-15km).
- 4 – CLOUD: pixels with cloud; include stratus and cumulus clouds that are on the low cloud level (from 0-0.2km to 2km).
- 5 – UNDEFINED: pixels that the labeler is not sure which class they belong to.

The dataset was labelled using Computer Vision Annotation Tool (CVAT) and Segments.ai. With the possibility of integrating active learning process in Segments.ai, the labelling was performed semi-automatically.

The dataset limitations must be considered: the data is covering only terrestrial region and does not include water areas; the dataset is not presented in winter conditions; the dataset represent summer conditions, therefore September and October contain only test products used for validation. Current subscenes do not have georeferencing, however, we are working towards including them.

2. Dataset Description

The provided dataset has the following structure:

```
DATASET
└── L1C
    ├── <month>
    │   ├── <L1C product>
    │   │   ├── <tile_x_y>
    │   │   │   └── <NetCDF.nc>
    ...
    └── estonia
        ├── <L1C product>
        │   ├── <tile_x_y>
        │   │   └── <NetCDF.nc>
```

L1C sub-folder contains 7 months ranging from April to October.

estonia contains 6 completely labeled products of Estonia for months of May, July, August.

Each **<month>** contains a set of selected L1C products that represent different cloud types from different geographical locations (Fig. 1).

Each **<L1C product>** contains sub-tiles 512x512 pixels in size obtained through cm-vsm and stored in NetCDF4 format in respective folders. The number of tiles varies depending on a product, but generally there are around 10 tiles per each L1C product.

Each **<NetCDF.nc>** includes the following series of bands: “B01” (443 nm), "B02" (490 nm), "B03" (560 nm), "B04" (665 nm), "B05" (705 nm), "B06" (740 nm), “B07” (783 nm), "B08" (842 nm), "B8A" (865 nm), “B09” (940 nm), “B10” (1375 nm), "B11" (1610 nm), "B12" (2190 nm), “Label”. The filename provide information about subscene coordinates which can be extracted by multiplying coordinates by 512 pixels (Figure 2). 512 x 512 pixles NetCDF sub-tiles are generated in tool developed by KappaZeta that is available by link: https://github.com/kappazeta/cm-vsm.

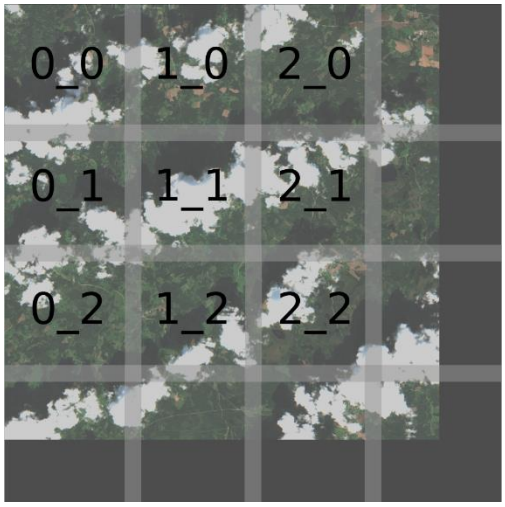


Figure 2. Illustrative images of how full Sentinel-2 is divided. Every NetCDF file in dataset has tile_x_y.nc name, where x represents the number for x axis and y the number of y axis and the exact coordinates from full images can be obtained by multiplying x and y by 512 pixels.

A set of test products, in addition, include FMC (Fmask classification map), SS2C (Sinergise S2Cloudless classification map), MAJAC (CNES MAJA cloud classification map). Note that MAJAC is available for a very limited number of products, as they should be located in 60°North and 56°South latitudes The

list of products that contain Fmask, S2Cloudless and MAJA for test comparison is in Appendix A.

The features are resampled to the same 10 m resolution with Sinc Infinite Impulse Response (IIR) filter that is windowed with a Blackman filter.

Acknowledgements

The data were collected, processed, and checked as a part of "KappaMask: AI-based Cloudmask Processor for Sentinel-2" project.

We thank CVAT and segments.ai teams for providing wonderful annotation tools that were actively to prepare the dataset.

In the end, we thank European Space Agency (ESA) for supporting, advising, and funding the project.

APPENDIX A: List of test products with additional masks (Fmask, S2Cloudless, MAJA*)

S2B_MSIL1C_20200426T101549_N0209_R065_T33VWF_20200426T131809
S2A_MSIL1C_20200415T100031_N0209_R122_T33UWT_20200415T121308
S2A_MSIL1C_20200413T092031_N0209_R093_T35ULT_20200413T111937
S2A_MSIL1C_20200503T092031_N0209_R093_T35UMQ_20200503T105308
S2B_MSIL1C_20200510T100029_N0209_R122_T33UWR_20200515T004952
S2A_MSIL1C_20200509T094041_N0209_R036_T34UEV_20200509T101545
S2B_MSIL1C_20200603T094029_N0209_R036_T35ULA_20200603T124101
S2B_MSIL1C_20200615T101559_N0209_R065_T33VWK_20200615T124248
S2A_MSIL1C_20200627T101031_N0209_R022_T33UWV_20200627T111749
S2B_MSIL1C_20200711T103629_N0209_R008_T32VNM_20200711T124043
S2A_MSIL1C_20200716T104031_N0209_R008_T33VVH_20200717T124424
S2A_MSIL1C_20200719T090601_N0209_R050_T36VVM_20200719T104105
S2A_MSIL1C_20200813T114401_N0209_R123_T29UNA_20200813T121202
S2B_MSIL1C_20200824T101559_N0209_R065_T32UQA_20200824T131033
S2B_MSIL1C_20200812T112119_N0209_R037_T30UVC_20200812T123206
S2B_MSIL1C_20200905T092029_N0209_R093_T35ULR_20200905T103628
S2B_MSIL1C_20200923T101649_N0209_R065_T33UUT_20200923T130930
S2A_MSIL1C_20200924T104031_N0209_R008_T31UFS_20200924T143955
S2B_MSIL1C_20201016T102929_N0209_R108_T32UMD_20201016T124151
S2B_MSIL1C_20201001T094039_N0209_R036_T35VMD_20201001T115620
S2B_MSIL1C_20201001T094039_N0209_R036_T35VMF_20201001T115620

*MAJA is available for a limited number of products