Radiant MLHub
EARTH IMAGERY FOR IMPACT

Dataset Documentation

Dataset Name: Sentinel-2 Cloud Cover Segmentation Dataset

Description

In many uses of multispectral satellite imagery, clouds obscure what we really care about - for example, tracking wildfires, mapping deforestation, or monitoring crop health. Being able to more accurately remove clouds from satellite images filters out interference, unlocking the potential of a vast range of use cases. With this goal in mind, this training dataset was generated as part of crowdsourcing competition, and later on was validated using a team of expert annotators. The dataset consists of Sentinel-2 satellite imagery and corresponding cloudy labels stored as GeoTiffs. There are 22,728 chips in the training data, collected between 2018 and 2020.

Citation

Radiant Earth Foundation. (2022). Sentinel-2 Cloud Cover Segmentation Dataset (Version 1) [Data set]. Radiant MLHub. https://doi.org/10.34911/RDNT.HFQ6M7

License

Creator

Radiant Earth Foundation

Contact

ml@radiant.earth

Tutorials

How to use deep learning, Pytorch Lightning, and the Planetary Computer to predict cloud cover in satellite imagery, Katie Wetstone

Location and boundaries

Overall Location Method

- □ Ground collection only
- □ Ground collection with boundary drawn using imagery
- □ Ground collection with spatial buffer added
- ☑ Boundary drawn from imagery
- □ Other _______________
- □ Unknown

Imagery Annotation methods

- □ Boundaries drawn based on a single ground point captured
- □ Boundaries drawn/edited based on multiple ground points captured

- □ Buffer validated from ground point captured
- ☑ Boundary drawn without ground reference data (Include description of methods in Appendix A)
- □ Pixels annotated without ground reference data (Include description of methods in Appendix A)
- □ Unknown

Boundary inclusion

- ☑ Captured polygon includes the entire field/area
- □ Captured polygon includes only a sample of the field/area
- □ N/A

Classification

Classification Type

- □ Land cover
- □ Crop type
- ☑ Other: Cloudy pixels

Classes/fields used

The only class identified in this dataset is cloud which contains all types of clouds (shallow or deep) and excludes cloud shadows.

Data Properties

Property name	Property Description	Parameters/Allowed responses
{chip_id}.tif / Band 1	The label tifs have a single band of data type Byte, indicating whether the pixel has clouds.	(0 = no clouds, 1 = clouds, 255 = nodata)

Appendix A: Describe how boundaries and classes were determined without ground reference data

The cloudy pixels for this dataset were generated and validated in two phases.

Phase 1: an open crowdsourcing competition was designed in which participants drew polygons on Sentinel-2 scenes to identify cloudy areas. In this phase, there was no validation of the labels and the purpose was to get the most number of labels. The scoring mechanism for the competition, however, encouraged participants to emphasize the detailed edges of the cloudy areas. Details are explained in this blog post.

Phase 2: The resulting labels from phase 1 were sent to an expert annotation group to validate, and if needed, correct the labels. The annotation team reviewed all the chips in the dataset and corrected any polygon that didn't correctly capture the cloudy areas, or added new polygons for cloudy areas that were missing.