Radiant MLHub
EARTH IMAGERY FOR IMPACT

Dataset Documentation

Name:
LandCoverNet South America: A Geographically Diverse Land Cover Classification Training Dataset

Publication Date:
March 31, 2022

Version:
1.0

Description:
LandCoverNet is a global annual land cover classification training dataset with labels for the multi-spectral satellite imagery from Sentinel-1, Sentinel-2 and Landsat-8 missions in 2018. LandCoverNet South America contains data across South America, which accounts for ~13% of the global dataset. Each pixel is identified as one of the seven land cover classes based on its annual time series. These classes are water, natural bare ground, artificial bare ground, woody vegetation, cultivated vegetation, (semi) natural vegetation, and permanent snow/ice.

There are a total of 1200 image chips of 256 x 256 pixels in LandCoverNet South America V1.0 spanning 40 tiles of Sentinel-2. Each image chip contains temporal observations from the following satellite products with an annual class label, all stored in raster format (GeoTIFF files):

- Sentinel-1 ground range distance (GRD) with radiometric calibration and orthorectification at 10m spatial resolution
- Sentinel-2 surface reflectance product (L2A) at 10m spatial resolution
- Landsat-8 surface reflectance product from Collection 2 Level-2

Radiant Earth Foundation designed and generated this dataset with a grant from Schmidt Futures with additional support from NASA ACCESS, Microsoft AI for Earth and in kind technology support from Sinergise.

Methodology:
To generate an annual land cover class label for each pixel several steps are taken, which are explained below. The specific details of each step are described in Alemohammad

and Booth, 2020. Labels are generated by looking at Sentinel-2 scenes. Sentinel-1 and Landsat 8 scenes are then added to the dataset.

- *Tile selection:* A representative set of Sentinel-2 tiles are selected to capture the diversity of global land cover classes (using MODIS MCD12Q1 V 5.1 as a guide). The number of tiles in each continent is proportional to the area of the continent.
- *Chip selection:* 30 chips of 256 x 256 pixels are selected in each tile to capture the diversity of different land cover classes within the tile.
- *Scene selection:* 24 scenes of Sentinel-2 are selected for each tile throughout 2018. These scenes are selected in a way to ensure there is at least one measurement in each calendar month sorted by cloud cover percentage.
- *Guess label:* A Random Forest (RF) model was generated using the 24 scenes at each tile to predict the land cover class at each 10 m pixel of Sentinel-2. GlobeLand30 land cover product was used as training data in this step.
- *Human validation:* A group of trained users were asked to validate or if needed change the guess label predicted by the RF model for each pixel. Each user would examine the 24 scenes throughout 2018, and see the predicted guess label. In case of any misclassification, they would correct the label and resubmit it. Users were also provided with high resolution imagery from Google basemap as auxiliary data, but in case of disagreement between Sentinel-2 time-series and Google basemap due to differences in their acquisition year, Sentinel-2 observations were used as the source of truth.
- *Consensus label:* Human interpretation error is unavoidable when labeling satellite imagery at 10 m spatial resolution. Therefore, each image chip was validated by three independent users. The accuracy of each user was assessed using chips that were separately labeled by experts from Radiant Earth's team. To generate the consensus label for each pixel a Bayesian model averaging approach was implemented taking into account the accuracy of each user. The resulting labels are accompanied by a "consensus score" between 0 and 100 which indicates the degree of agreement among the three users.

  In this step, each image chip was broken down into non-overlapping blocks of 32 x 32 pixels (a total of 64 blocks for each chip) to facilitate the label validation process. The resulting consensus labels for each 32 x 32 tasks are then mosaicked together to generate the 256 x 256 image chips. This might have caused artifacts in some of the chips where there is a sudden change in the label from one 32 x 32 block to the next one. It is recommended that users combine the consensus score layer for each chip with the label when using them for training or validation tasks.

- *Harmonizing spatial resolution of input data:* For each image chip, the corresponding time series of Sentinel-1 and Landsat 8 scenes throughout 2018 is retrieved and clipped to the bounding box of the chip. The grid from Sentinel-2 10m spatial resolution bands are used as target grid for source images. Therefore, Sentinel-2 bands which have a coarser resolution are mapped at 10 m using nearest neighbor interpolation. Sentinel-1 scenes come at 10m spatial resolution, but their grid is different from Sentinel-2; therefore, a nearest neighbor interpolation was applied to match their grids. Finally, Landsat 8 scenes which come at 30m spatial resolution were mapped to 10m spatial resolution of Sentinel-2 data using nearest neighbor interpolation.
- *Publication:* The final labeled dataset is published along with all the scenes from Sentinel-2, Sentinel-1 and Landsat products for 2018. For each scene of Sentinel-2, a cloud probability layer and the scene classification layer which are produced by Sen2Cor atmospheric correction package are included in addition to all the multi-spectral bands at 10 m.

Class Definitions:

Land cover classes are defined based on a hierarchical schema that was developed at an expert working group workshop hosted by Radiant Earth Foundation in June 2018. Based on the recommendations from the workshop, the following schema is used:

Level 1	Level 2	Level 3	Value	Color
Bare *(max veg/yr < 10%)*	Water *(max water/yr > 90%)*	Water (Permanent)	1	#0000ff
	Bare Ground	Artificial	2	#888888
		Natural	3	#d1a46d
	Snow/Ice *(max snow or ice/yr > 90%)*	Snow/Ice (Permanent)	4	#f5f5ff
Vegetated *(max veg/yr >= 10%)*	Woody	Woody	5	#d64c2b
	Non-Woody	Cultivated	6	#186818
		(Semi) Natural	7	#00ff00

The process to identify the land cover class for a pixel starts with examining the time-series of 24 scenes and calculating the percentage of times that the pixel is vegetated among the cloud free observations.

If the pixel is classified as Vegetated (max veg/yr >= 10%), then the pixel is classified as either Woody, Cultivated or (Semi) Natural vegetation. To decide between the three vegetation classes the annual time-series and in some cases the high resolution Google basemap (particularly for identifying woody vegetation) are used.

If the pixel is classified as Bare (max veg/yr < 10%), then the pixel is classified as Water (max water/yr >=90%), Snow/Ice (max snow or ice/yr >=90%), or Bare Ground otherwise. In the case of Bare Ground, and if needed, Google high resolution basemap is used to distinguish Artificial and Natural Bare Ground classes.

Note regarding Cultivated Vegetation class: a pixel is identified as Cultivated if it is planted in the year 2018, otherwise it is classified as Artificial Bare Ground in absence of any natural vegetation.

Coordinate Reference System:
Image chips are stored in UTM/WGS84 projection consistent with Sentinel-2 L2A projection and grid.

File Name Structure:
Each image chip is uniquely identified by a combination of a five character Sentinel-2 tile ID (e.g. 36RVP) and a two digit chip ID (range from 00 to 29). Source images for each chip are grouped by their observation date. The label and consensus score are provided in one GeoTiff file (band 0 is label and band 1 is consensus score). A csv file is provided for each chip containing dates of the 24 scenes that were used to identify the annual label of each pixel.

The file name structure for each chip is described below.

Source Imagery:

<XXXXX>/<NN>/<PP>/<XXXXX>_<NN>_<YYYYMMDD>/<XXXXX>_<NN>_<YYYYMMDD>_<ZZZ>_10m.tif

Label:

<XXXXX>/<NN>/<XXXXX>_<NN>_LC_10m.tif

Labeling Imagery Dates:

<XXXXX>_<NN>_labeling_dates.csv

In which:

<PP>: Platform for source imagery (S2, S1 or L8)

<XXXXX>: Sentinel-2 tile ID

<NN>: Chip ID (00 to 29)

<YYYYMMDD>: Observation date of source imagery scene

<ZZZ>: Band ID of source imagery scene. For Sentinel-2 they contain B01, B02, B03, B04, B05, B06, B07, B08, B8A, B09, B11, B12, CLD, and SCL (CLD represents cloud probability, SCL is the scene classification layer.) For Sentinel-1, there are two combinations VV and VH. For Landsat 8 there are 7 bands: B01, B02, B03, B04, B05, B06, B07.

Spatial Extent:
The geographical coverage of the data is the continent of South America. The latitude and longitude of the corresponding bounding box is:

44.28688968799801 S, 78.20135456849277 W
9.949301486520056 N, 36.193468268263175 E

Temporal Extent:
The land cover labels are based on the time-series of source imagery in 2018, and the temporal extent is from 2018/01/01 to 2018/12/31. For simplicity, labels are timestamped to 2018/07/01 in the catalog.

Citation:
Radiant Earth Foundation (2022) "LandCoverNet South America: A Geographically Diverse Land Cover Classification Training Dataset", Version 1.0, Radiant MLHub. https://doi.org/10.34911/rdnt.6a27yv

Contact:
Radiant Earth Foundation
support@radiant.earth