TransitionZero

TZ-SAM: Solar Asset Mapper Data Summary

29th May 2024

TransitionZero's Solar Asset Mapper (TZ-SAM) is a global, satellite-derived dataset of utility-scale solar farms generated with a combination of machine learning and human annotation. Our Q1 2024 dataset contains the location and shape of 63,616 assets, along with estimated capacities. We estimate the construction date for over 80% of these assets. The dataset contains over 19,100 square kilometres of solar farms across 183 countries, with a total estimated capacity of 711 GW.

1. Dataset Description

We publish six files:

- ***analysis_polygons.gpkg***: our "analysis-ready" dataset containing geometries, capacity estimates and construction date estimates.
- ***analysis_polygons.csv***: a version of *analysis_polygons.gpkg* containing a central latitude and longitude in place of a geometry, to allow parsing without geospatial software.
- ***sources.csv***: a table mapping the IDs of our analysis-ready dataset to the raw geometries that make them up.
- ***raw_polygons.gpkg***: the raw geometries used to compose *analysis_polygons.gpkg*.
- ***TZ Solar Asset Mapper Q1 2024.xlsx***: an Excel formatted version of the *analysis_polygons.csv* file.
- ***tz-sam_scientific_data.pdf***: A pre-print aricle that explains the methodology in detail.

1.1 Analysis-level datasets

Our analysis-level dataset comprises our most complete view of global asset-level solar installations, incorporating our own detections as well as known solar farm geometries from other datasets.

The geospatial dataset contains the following fields:

- **id**: unique ID for the asset
- **geometry**: Polygon or MultiPolygon defining the asset
- **capacity_mw**: estimated capacity of the asset in megawatts
- **constructed_before**: upper bound for construction date (estimated date of the image in which the solar plant was first seen in a constructed state)
- **constructed_after**: lower bound for construction date (estimated date of the image in which construction began for the solar plant)

The CSV version replaces the Geometry column with:

- **latitude**: the latitude of the centroid of the asset
- **longitude**: the longitude of the centroid of the asset
- **country**: administrative country name

1.2 Raw datasets and sources

The analysis-level datasets hide some complexity in the underlying data that we expose in the *raw_polygons* and sources file.

- We produce new sets of polygons for each run. Often these overlap, sometimes in complicated ways.
- We cluster together overlapping and nearby geometries from both our detections and external sources. Currently these sources are:
  - Large solar farms scraped from OpenStreetMap (OSM)
  - Validated geometries from Kruitwagen et. al., A global inventory of solar photovoltaic generating units.

Each cluster comprises one row in the analysis-level dataset. In order to enable tracking raw detections from run to run, as well as to provide detailed sourcing information, we provide all of these raw polygons, along with a source file that lists all of the raw polygons contained in each analysis-level polygon.

raw_polygons.gpkg contains the following fields:

- **id**: ID of the raw source polygon
- **geometry:** Polygon or MultiPolygon defining the asset
- **source**: either "solar asset mapper", "osm" or "2019_global_pv".
- **acquisition_date**: for solar asset mapper polygons, this is the date of the inference run that produced the polygon; for OSM polygons it is the date that the polygon was scraped from OSM; for 2019_global_pv it is 2019-01-01, the approximate detection date of that dataset.

Sources.csv contains the following fields:

- **cluster_id**: ID of the corresponding item in the analysis-level dataset
- **source_id**: ID of the raw source polygon
- **source**: either "solar asset mapper", "osm" or "2019_global_pv".
- **acquisition_date**: for solar asset mapper polygons, this is the date of the inference run that produced the polygon; for OSM polygons it is the date that the polygon was scraped from OSM; for 2019_global_pv it is 2019-01-01, the approximate detection date of that dataset.

1.3 Caveats and limitations

False positives

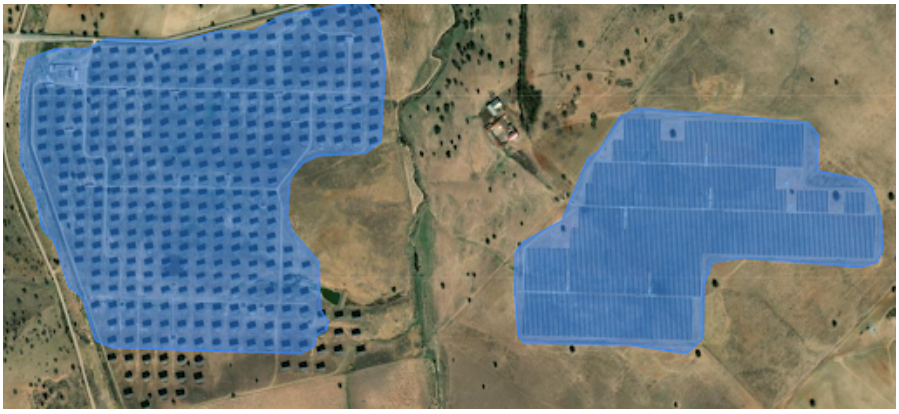
While we have made every effort to remove false positives from the published dataset, some will remain due to the difficulty of manually validating detections in 10-metre satellite imagery. To estimate false positive prevalence throughout the data a subset of approximately 2000 detections were selected at random from our positively labelled solar assets. Each of these were validated through a higher degree of scrutiny utilising high-resolution imagery. This analysis yielded an expected rate of false positives of around 1%.

Plant shapes

Our plant outlines are not perfect. They will occasionally be much smaller or larger than the underlying plant. Our tests show that on average, these effects average out.

Capacity estimates

Our capacity estimation model should produce relatively unbiased country-level aggregates, since it is trained to learn the typical ground coverage ratio of plants by country. The model has no way to distinguish between a very dense and a very sparse (e.g. dual-axis-tracking) plant in the same country. Plants with unusually high or low ground coverage ratios will not have accurate capacity estimates.



Left: a dual axis facility. Right: a static facility. The static facility has a notably higher ground coverage ratio (GCR) and therefore greater capacity. This is not directly captured however. GCR - and by extension capacity - estimates for both of these facilities are based on the size of the facility and country of origin.

Construction date estimates

We are not able to directly estimate the construction date of a plant. We estimate an upper bound (the date of the image in which the plant was first seen in a constructed state) and a lower bound (the date of the image in which the plant was last seen in an unconstructed state). For plants that were constructed before 2017, we produce only an upper bound. This is for two reasons. First, Sentinel-2A launched in 2015, resulting in no imagery to use before then. Second, the imagery available before the launch of Sentinel-2B in 2017 was sparse and therefore deemed unreliable in our analysis.

We leave it to consumers of the data to interpret these bounds and/or estimate likely grid connection dates.

2. Attribution

Attribution to TransitionZero is required per the TZ-SAM Terms of Use. The TZ-SAM data and associated metadata has been made available via TransitionZero under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0). This means that anyone is free to use TZ-SAM data in any format for non-commercial purposes only.

You must also clearly indicate if you have made any changes to the TZ-SAM dataset and what these are. Please refer to the suggested citation formats:

- "TransitionZero Solar Asset Mapper, TransitionZero, May 2024 release."
- "TZ-SAM, TransitionZero, May 2024 release."
- "TransitionZero (2024) Solar Asset Mapper."