TZ-SAM: Solar Asset Mapper Data Summary

TransitionZero's Solar Asset Mapper (TZ-SAM) is a global, satellite-derived dataset of utility-scale solar assets that has been generated using a combination of machine learning methods and human annotation. The Data Summary document describes the included data files, references the source data, and summarizes caveats and limitations of the data.

1. Dataset description

We provide four files within the TZ-SAM data package:

- ***analysis_polygons.csv***: our "analysis-ready" dataset containing a central latitude and longitude in place of a geometry, to allow parsing without geospatial software.
- ***analysis_polygons.gpkg***: our geospatial "analysis-ready" dataset containing geometries, capacity estimates and construction date estimates.
- ***raw_polygons.csv***: the raw polygons' sources and a mapping to the clustered polygons in *analysis_polygons.csv*.
- ***raw_polygons.gpkg***: the raw geometries, their sources, and a mapping to the clustered polygons in *analysis_polygons.gpkg*.

1.1 Analysis-level dataset

The analysis-level dataset provides our most complete, asset-level view of solar installations globally. It is a combination of TransitionZero's novel detections and known solar farm geometries from other datasets.

The geospatial *analysis_polygons.gpkg* dataset contains the following fields:

- **cluster_id**: unique ID for the clustered asset
- **capacity_mw**: estimated capacity of the asset in megawatts (MW)
- **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)
- **geometry**: Polygon or MultiPolygon defining the asset

The *analysis_polygons.csv* dataset replaces the geometry column with:

- **latitude**: the latitude of the centroid of the asset
- **longitude**: the longitude of the centroid of the asset

and adds a column for:

- **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* files.

- 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:
  - **sam-<year>-<quarter>**: Previously detected, novel solar assets from TransitionZero
  - **gpv-model**: Validated geometries from *Kruitwagen et. al. (2021): A global inventory of solar photovoltaic generating units*.
  - **china-annotator, china-model**: Validated geometries from *Zhang et al. (2020): The dataset of photovoltaic power plant distribution in China by 2020*.
  - **india-model**: Validated geometries from *Ortiz et al. (2022): An Artificial Intelligence Dataset for Solar Energy Locations in India*.
  - **uspvdb-annotator**: Validated geometries from *Fujita, K.S. et al. (v2, 2024): United States Large-Scale Solar Photovoltaic Database*.

Each cluster comprises one row in the analysis-level dataset. To enable the tracking of raw detections from run to run, and to provide detailed sourcing information, we provide all of these raw polygons and a mapping to the clusters they belong to within the analysis-level data.

raw_polygons.gpkg 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**: a key representing the polygon data source
- **geometry:** Polygon or MultiPolygon defining the asset
- **source_date**: The approximate date that a solar asset was detected by a source

The *raw_polygons.csv* dataset replaces the geometry column with:

- **latitude**: the latitude of the centroid of the asset
- **longitude**: the longitude of the centroid of the asset

and adds a column for:

- **country**: administrative country name

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 within 10-metre-resolution satellite imagery. We separately estimate the false positive rate of each of our external data sources and of our own validated detections. To do this, we sample 1000 images at random from each source and manually validate them using high-resolution imagery. As a result of this analysis, we expect the total prevalence of false positives from each source - and of the dataset as a whole - to be less than 0.5%.

Plant shapes

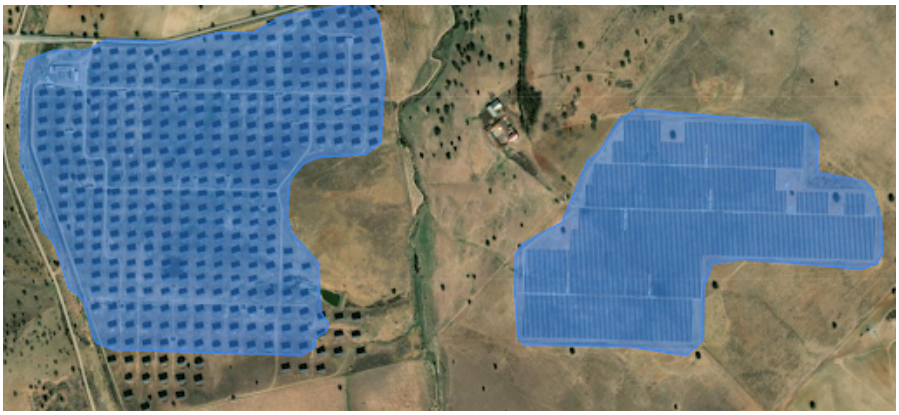
Our plant outlines are not perfect. They will occasionally be much smaller or larger than the underlying plant. Our tests show that on average, these effects average out.

Capacity estimates

We train a linear model to predict the power density of a detection based on its size and location. As of the Q3 2024 release, we have updated this model to include the estimated construction date of the plant. This model accounts for the following effects that we observe in the data:

- Typical plant layouts and panel packing density vary significantly from country to country
- Larger plants tend to be constructed more sparsely than smaller ones
- Commercial solar cell efficiencies increase over time

This model should produce relatively unbiased country-level aggregates, as it is trained to learn the typical ground coverage ratio of plants by country. However, 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 cannot 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: Sentinel-2A launched in 2015, meaning no imagery is available to use before that year; the imagery available before the launch of Sentinel-2B in 2017 is sparse and has been deemed unreliable for the purpose of 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 change are. Please refer to the suggested citation formats:

- "TransitionZero Solar Asset Mapper, TransitionZero, <Month Year> release."
- "TZ-SAM, TransitionZero, <Month, Year> release."
- "TransitionZero (<Year>) Solar Asset Mapper."