

Estimating distributed solar in Pakistan

Technical methodology

Updated: 20 October 2025

Introduction

To estimate the amount of small-scale solar adoption in Pakistan, such as rooftop and ground-mounted solar, high-resolution imagery is required. High-resolution imagery for detecting small-scale solar assets requires a resolution of at least 0.3m, which comes at a cost. Obtaining this imagery for a country the size of Pakistan would be too expensive for this project, let alone considering the future expansion to other regions and continents. In other words, achieving the global coverage that we would eventually want would require a lot of investment.

If we can't use machine learning, what are the other options? Our approach to this problem was to use sampling and extrapolation techniques, enabling us to make estimates for rooftop solar capacities, without the need for detecting them on an individual basis. This meant that we could obtain satellite imagery for a much smaller area, and therefore cost, whilst still being able to make accurate estimates for rooftop solar across the country.

Choosing an approach for rooftop solar

In order to carry out this method, we needed a country-wide dataset which we believed to be correlated with the distribution of rooftop solar. For example, we could calculate the rooftop solar per capita within a sample region, before extrapolating to the whole country. Similarly, we could use building count, or building area, with the assumption that higher levels of these things would correlate with higher levels of rooftop solar.

We settled on using buildings data, as the available datasets for population were not suitable for our use case. In particular, they aren't at the level of granularity required for the typical images we would be using ($\sim 0.1\text{km}^2$). This would introduce significant uncertainty to our calculations due to the assumptions being made, such as the population size within each image.

Therefore, we proposed the following methodology, using buildings data, for our Pakistan study:

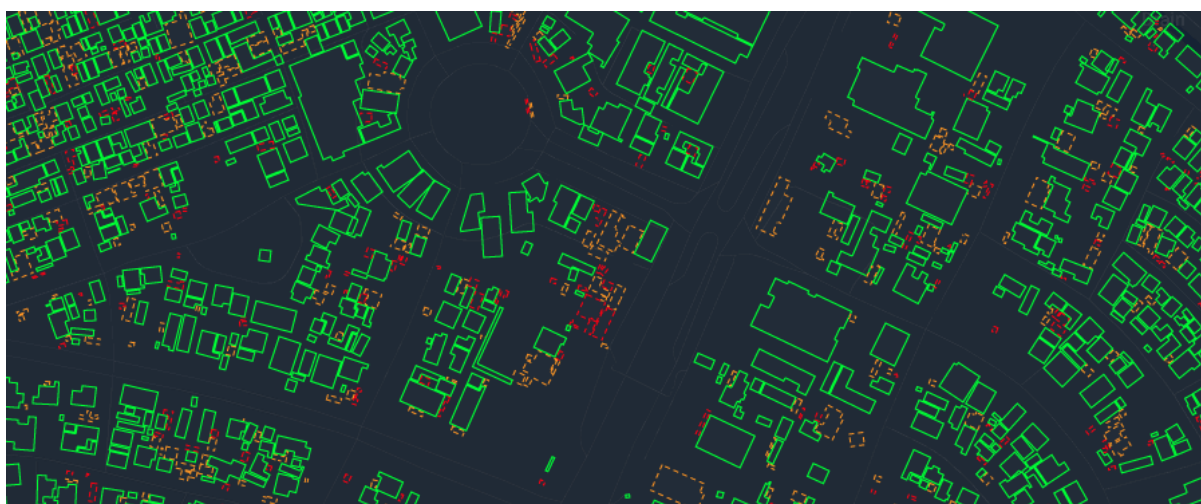
1. Sample 1000 locations throughout Pakistan
2. Obtain high resolution satellite imagery for each location

3. Use manual and machine learning based methods to label all the rooftop solar in each image
4. Establish estimates over our sample of rooftop solar coverage per building
5. Get estimates for the total rooftop areas in Pakistan
6. Extrapolate our rooftop solar estimates to a country and regional level

Throughout the project, we used the Open Buildings dataset from Google, in both our sampling and extrapolation methods. This dataset contains polygons for all of the building footprints within the countries that it covers (including Pakistan).

We used this data for both sampling (choosing locations that contain rooftops) and extrapolation (estimating the total rooftop area within the country or a region). In simple terms, we could then calculate the total estimated rooftop solar using:

$$\text{Total Rooftop Solar Area} = \text{Total Rooftop Area} \times \text{Estimated Rooftop Coverage}$$



Google's Open Buildings data in Lahore, Pakistan

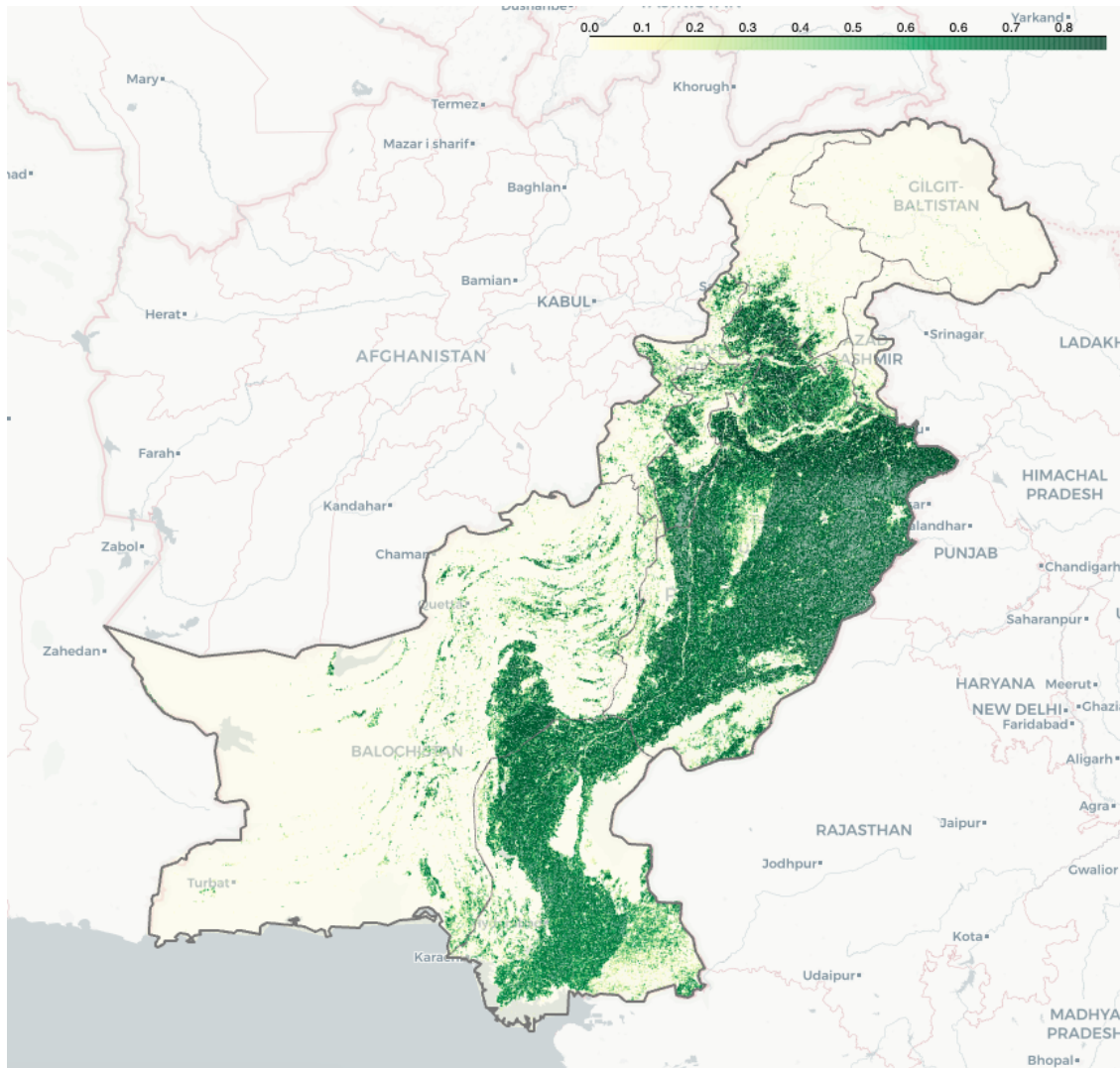
Ground-mounted solar

A unique challenge posed by Pakistan is the amount of ground-mounted solar installed throughout the country. The majority of these are believed to be in the agricultural sector, used for solar-based tube wells and water irrigation systems. It was also thought that the total capacity of these types of installations is large enough to contribute significantly to Pakistan's total, so it was important to be included in this study.



Solar powered tube wells in Pakistan. Credit: CGIAR

The difficulty with ground-mounted solar is that we were unable to use buildings data (or rooftop areas) in our extrapolation methods. From our research and discussion with PRIED, it was decided that the strongest correlation with ground-mounted solar installations would be agricultural landcover data. We decided to use data from Dynamic World, which contains data for “Crops” land cover, a strong proxy for agricultural land.



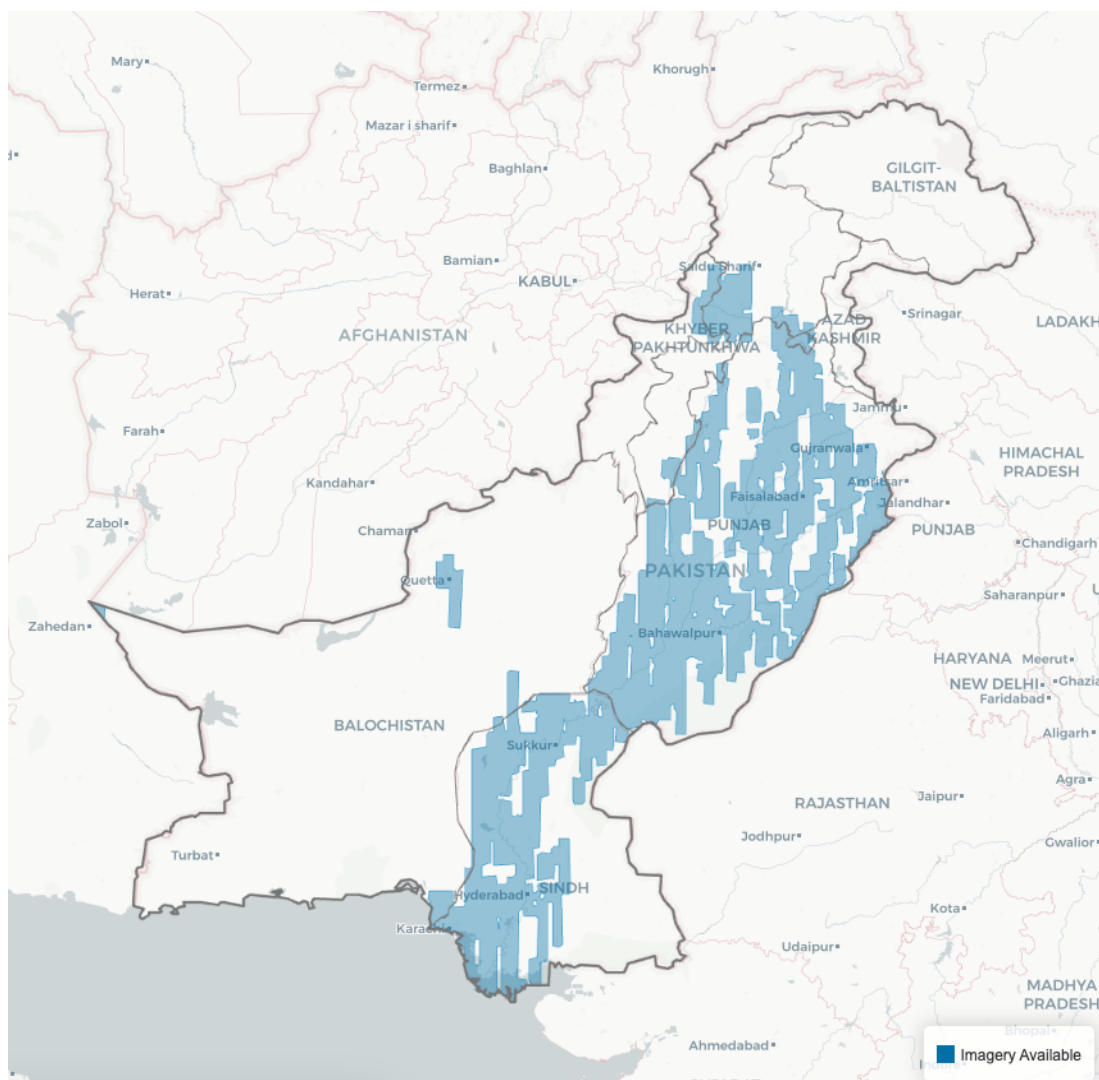
Aggregated agriculture landcover data from Dynamic World for Pakistan

The rest of the methodology follows the same process as for rooftop, with the main difference being in the final extrapolation. For example, it could be simplified to:

$$\text{Total Ground-mounted Solar Area} = \text{Total Agricultural Area} \times \text{Estimated Ground Coverage}$$

Sampling

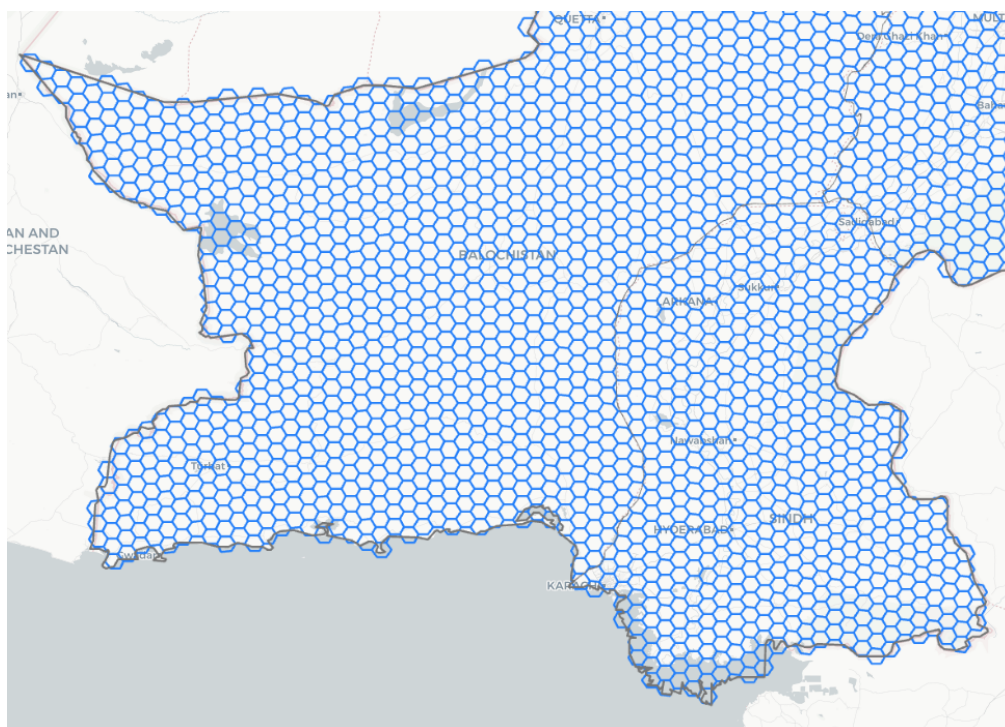
Firstly, our available sampling region was restricted due to the availability of recent (< 12 months old) imagery. We were using 0.3m resolution imagery from the Pléiades Neo (PNEO) satellite, which only covers the more populous regions of a country for images at the recency, and cost that we required.



Availability of 0.3m imagery from PNEO across Pakistan

Despite missing large fractions of the country, the most populous and inhabited regions of Pakistan are covered. We believed that the coverage was sufficient to enable a representative sample and an accurate study.

We chose to sample 1000 locations across Pakistan, for which we would obtain high-resolution satellite imagery, each covering an area of 0.1 km². To build this sample, we divided the country into H3 grid cells (of resolution 9) and then applied a series of selection criteria. H3 is a way of indexing geographies using a hexagonal grid. At resolution 9, this divides Pakistan into approximately 8 million cells of ~0.1km² each.



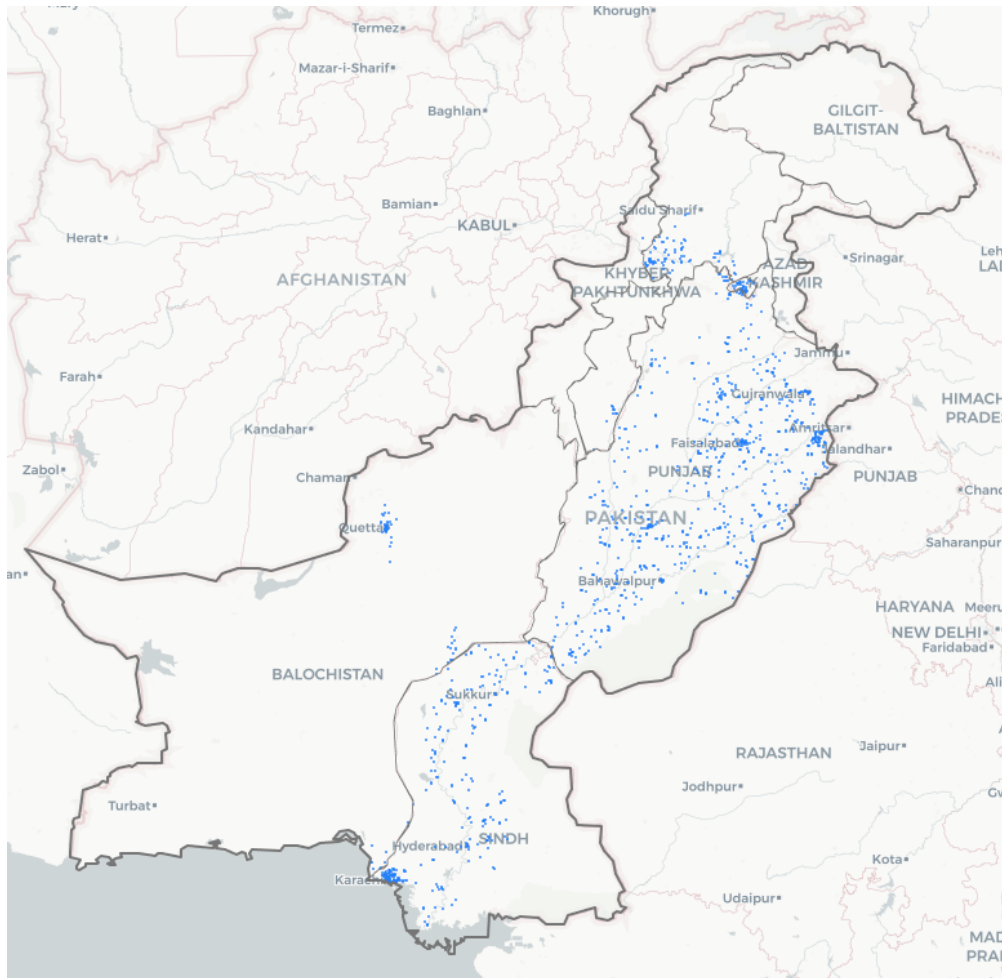
*Southern Pakistan divided into H3 cells of resolution 5 (~252 km² per cell).
For our study, we used resolution 9 (~0.1 km² per cell).*

First, we ensured that no two sampled cells were adjacent, in order to avoid overlap and redundancy. We also excluded any cells where no suitable imagery had been captured in the previous twelve months, ensuring that our analysis reflected the most recent conditions. Using the Open Buildings Dataset, we were able to exclude cells in remote areas that had fewer than three buildings and no agricultural land, as these had a high rate of being barren remote areas, which would lead to wasted imagery.

In order to ensure we covered enough agricultural areas, we sampled all cells with some agricultural coverage and at least 1 building. Additionally, we oversampled in Balochistan to ensure an adequate sample, due to its vast solar potential, and the limitations in image availability in this region. This was further complemented with the on-ground survey conducted by PRIED.

To distinguish between urban and non-urban areas, we calculated building density per cell (using Google's Open Buildings) and applied a 30 percent threshold: cells with more than 30 percent building coverage were categorized as urban, while all others were classified as non-urban. From these two categories, we drew an even number of samples, selecting locations at random to minimize bias. Splitting our sample in this way ensured that our sample contained both highly urbanized areas (such as cities) as well as more rural communities (such as towns, villages and agricultural land).

The graphics below show the distribution of our sample imagery and a breakdown by province.



Locations of ~1000 sample images

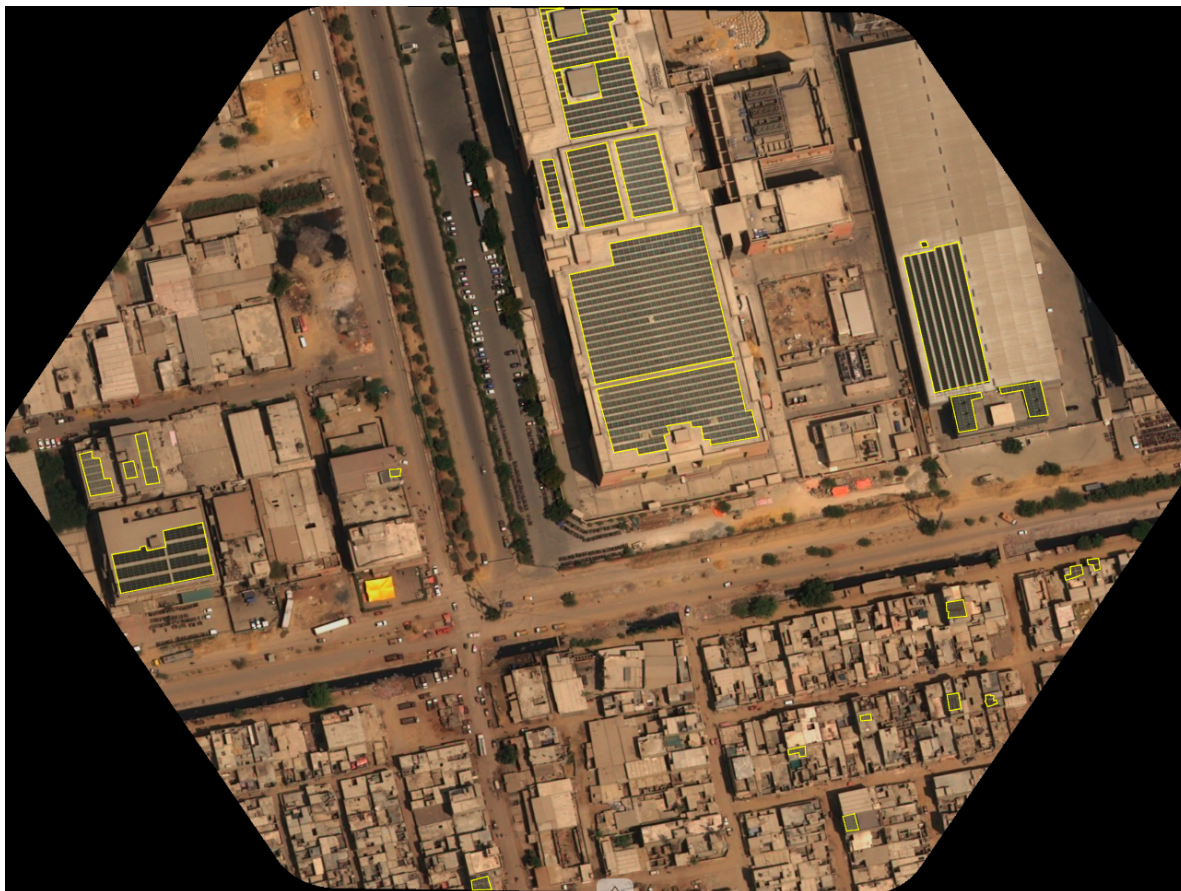
Province	No. of samples
Punjab	609
Sindh	264
K.P.	79
Balochistan	59
I.C.T.	27
Total	1038

Total sampling numbers by province

Labelling

For each image in our sample, we used a combination of machine learning and manual labeling techniques for outlining each solar installation present. All machine learning based labels were manually inspected to ensure accuracy and completeness. The labels were

consolidated into a labeling tool, CVAT, and further classified into either rooftop and ground-mounted solar.



© Airbus DS (2025) | Example rooftop solar labels using CVAT

Analysis and extrapolation

In total, there were ~27,000 solar labels produced across our sample, which we analysed in two separate phases (rooftop and ground-mounted). For each, we produced an extrapolation model, based on the insights from our sample, that we could apply to the whole of Pakistan.

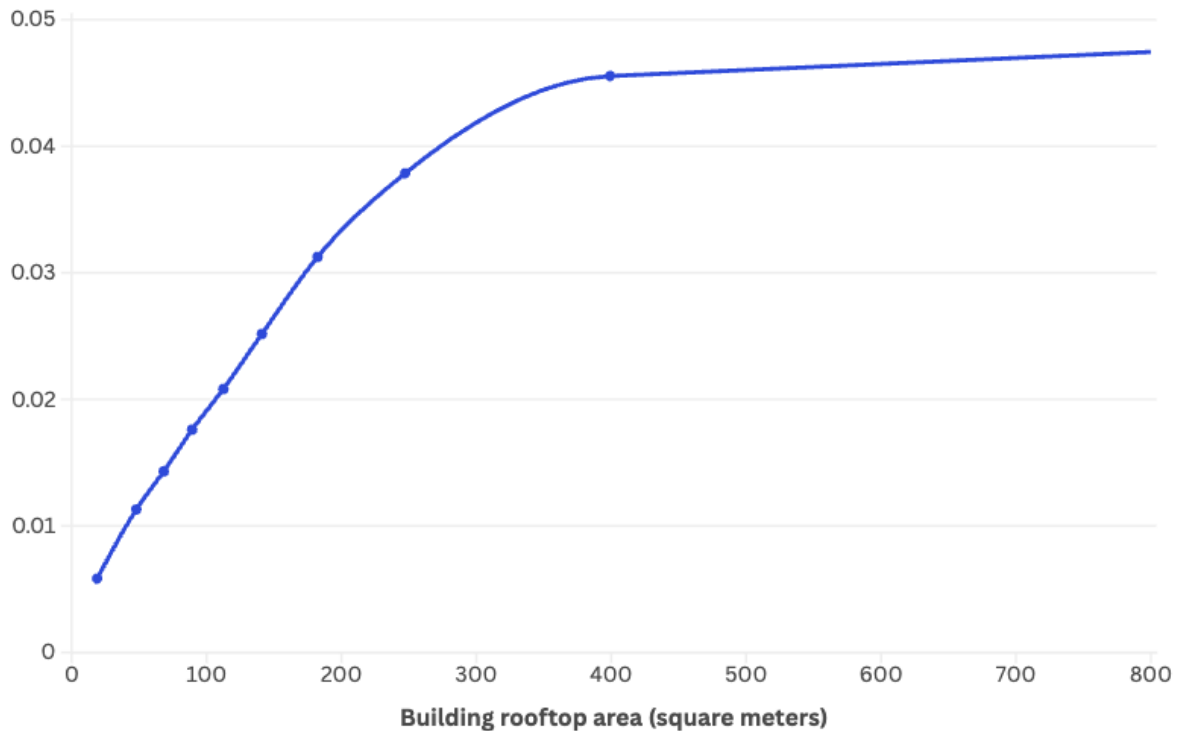
Rooftop analysis

For rooftop solar, our extrapolation model is based on building size. Looking at the data, we could clearly see a correlation between rooftop area and solar coverage. i.e. larger buildings are more likely to have a higher proportion of their rooftops covered by solar.

Bigger rooftops mean more solar

Solar coverage rises steadily with rooftop size before plateauing at around 400 square metres

Average solar coverage per square meter



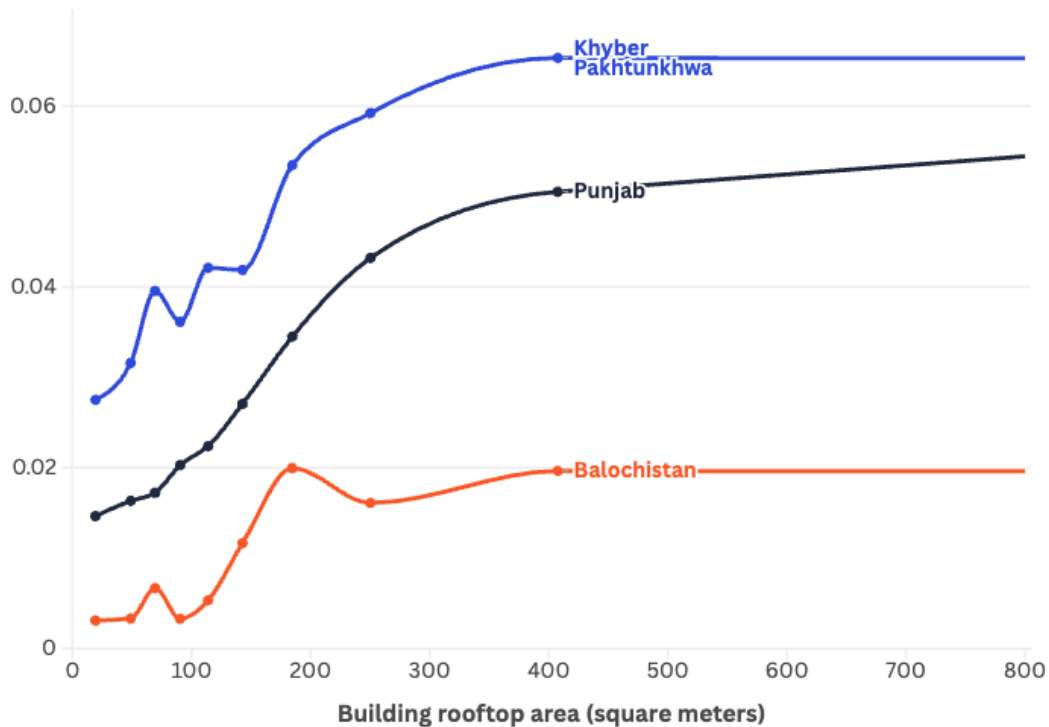
TransitionZero

Looking at this same relationship regionally, there are also clear differences in solar installation rates across the country. The plot below shows a subset of regions with the largest differences. Most notably, K.P. shows much higher rates of installation than elsewhere in Pakistan.

Rooftop solar coverage by building size across regions

Khyber Pakhtunkhwa shows the highest rooftop coverage, followed by Punjab and Balochistan

Average solar coverage per square meter



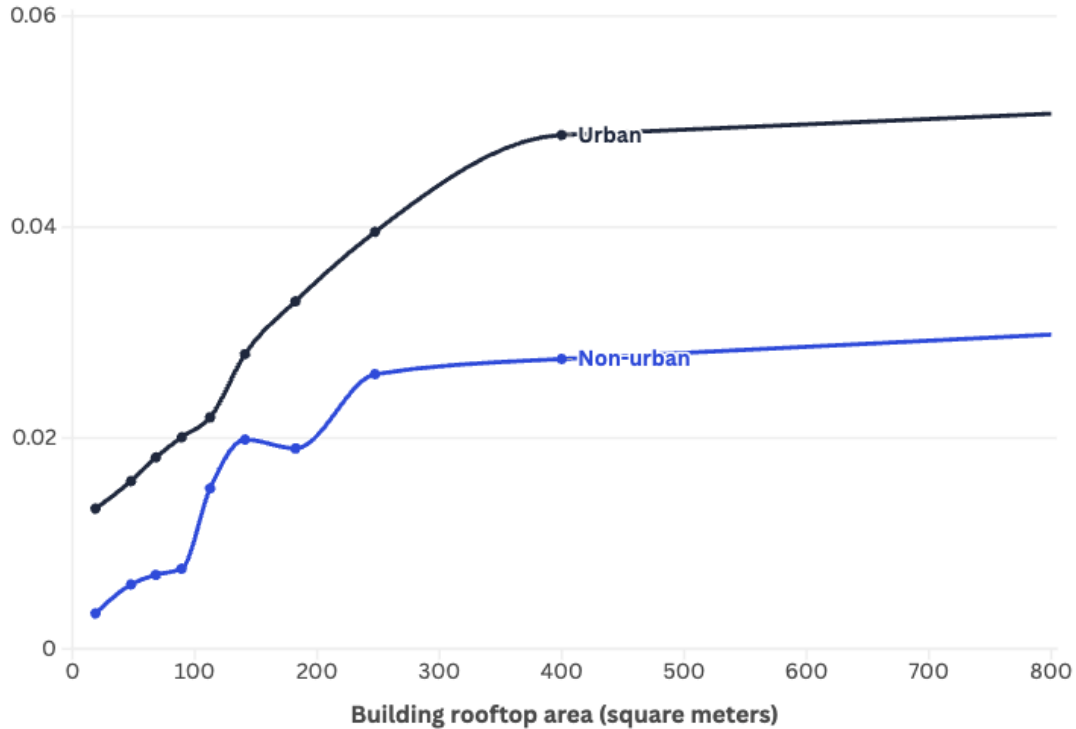
TransitionZero

If we look at the same data across different levels of building density (urban vs. non-urban) we see a trend of higher rates in urban areas (such as cities) compared to non-urban areas (such as smaller villages or rural communities). The plot below shows this relationship across the whole of Pakistan.

Urban rooftops achieve higher solar coverage than rural ones

Comparing solar coverage by rooftop size across Pakistan's urban and rural buildings

Average solar PV coverage per square meter



TransitionZero

Using these insights we applied a region-level building size model across all buildings in Pakistan, to produce an estimated solar coverage in m². We then applied a correction factor, to any buildings in 'non-urban' areas, to allow for the observed lower rates in these areas.

Finally, to convert from predicted solar m² to capacity, we applied the following factors:

- Panel efficiency
 - 21% for urban areas
 - 17% for non-urban and agricultural areas
- Panel tilt correction
 - 1.16 due to latitude of Pakistan and average tilt
- Label panel coverage
 - 0.75 to account for empty space within solar labels

Ground-mounted analysis

For our ground-mounted solar extrapolation we used a model based on crop landcover data per H3 cell. Using data from Dynamic World, we were able to measure the relationship between agricultural land and the rate of ground-mounted solar installations. When broken

down by province, it's clear that certain areas have higher adoption of solar than others in the agricultural sector.

Province	Solar m ² per unit of agricultural landcover
Balochistan	0.29
K.P.	0.25
Punjab	0.20
Sindh	0.13

We then applied this rate to all H3 cells throughout Pakistan, whilst also applying the assumption that any cells with zero buildings have zero solar.

We also applied a small correction for cells where there is zero crop coverage, to account for ground-mounted solar in urban areas.

Uncertainties

The major sources of uncertainty are listed below with a brief explanation for each. These uncertainties are used when calculating an error range in our results.

- Labelling confidence
 - During labelling, our confidence in objects being solar varied due to both object size and image quality. Smaller objects are more difficult to classify with certainty as being solar, whereas unclear imagery (due to atmospheric changes or light conditions) also made this a challenge in certain cases. This led to some of our solar labels being of a lower confidence than others.
- Panel efficiency
 - As noted above, we used estimated panel efficiencies in our extrapolation calculations. However, the exact value of these efficiencies is unknown and is likely to vary by region, location and age of panels.
- Sampling bias
 - Also mentioned above is the restriction of imagery availability from PNEO across Pakistan. This means that certain areas of the country are covered less comprehensively (or not at all) by our study, which introduces some uncertainty into the assumptions made during our extrapolation.
- External datasets
 - Both Open Buildings and Dynamic World are datasets which, although incredibly useful, are not perfect and come with their own levels of uncertainty and error. We use both of these in our extrapolation models, so this will contribute a certain level of error to our final results.
- Panel tilt
 - During our extrapolations we applied a correction factor to account for the fact that solar panels in our imagery are likely to be tilted for optimum efficiency.

This will vary throughout Pakistan, due to location, but also locally, due to user error and varying levels of knowledge in panel installation.

Data citations

- W. Sirko, S. Kashubin, M. Ritter, A. Annkah, Y.S.E. Bouchareb, Y. Dauphin, D. Keysers, M. Neumann, M. Cisse, J.A. Quinn. *Continental-scale building detection from high resolution satellite imagery*. [arXiv:2107.12283](https://arxiv.org/abs/2107.12283), 2021.
- Brown, C.F., Brumby, S.P., Guzder-Williams, B. *et al.* Dynamic World, Near real-time global 10 m land use land cover mapping. *Sci Data* 9, 251 (2022). <https://doi.org/10.1038/s41597-022-01307-4>