Final Technical Report (FTR)

Cover Page

a. Federal Agency	Department of Energy	
b. Award Number	DE-EE0009359	
c. Project Title	Machine-Learning-Based Mapping and Modeling of Solar Energy with Ultra-High Spatiotemporal Granularity	
d. Recipient Organization	Stanford University	
e. Project Period	Start: 8/1/2021	End: 7/31/2023 No-cost Extension: 8/1/2023 – 1/31/2024
f. Principal Investigator (PI)	Ram Rajagopal Associate Professor Email: ramr@stanford.edu Phone: (650) 725-4268	
g. Business Contact (BC)	Layton Dutton Contract and Grant Officer Email: laytonh@stanford.edu Phone: (650) 724-8889	
h. Certifying Official (if different from the PI or BC)	Name Title Email address Phone number	

Ram Rajagopal
Signature of Certifying Official

05/30/2024
Date

By signing this report, I certify to the best of my knowledge and belief that the report is true, complete, and accurate. I am aware that any false, fictitious, or fraudulent information, misrepresentations, half-truths, or the omission of any material fact, may subject me to criminal, civil or administrative penalties for fraud, false statements, false claims or otherwise. (U.S. Code Title 18, Section 1001, Section 287 and Title 31, Sections 3729-3730). I further understand and agree that the information contained in this report are material to Federal agency's funding decisions and I have any ongoing responsibility to promptly update the report within the time frames stated in the terms and conditions of the above referenced Award, to ensure that my responses remain accurate and complete.

1. Acknowledgement

This material is based upon work supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) Solar Energy Technologies Office (SETO) under the Fiscal Year 2020 Funding Program, Award Number DE-EE0009359.

2. Disclaimer

3. Executive Summary

Despite the rapid growth of solar energy, we still lack a dynamic, high-fidelity database that tracks the spatiotemporal variations of solar PVs and their associated infrastructures across different places at a spatially resolved scale. The absence of such data presents a barrier to various applications such as solar PV growth projection, solar energy integration, solar incentive design, and climate risk assessment. In this project, we aim to bridge this gap by developing AI-based algorithms to extract granular information about solar PV installations and their associated infrastructures (i.e., distribution grids) from widely available unstructured data like remote sensing images and street views. As a result, we have built the Solar Energy Atlas, a fine-grained, large-scale geospatial overlay of distributed solar PVs and distribution grids. On top of it, we have advanced the understanding of solar adoption and distribution grid vulnerability to climate-induced extremes. Our major contributions can be summarized as follow:

- By developing new AI algorithms, we have built the most comprehensive solar PV spatiotemporal database covering the entire US. This is the first time we obtained the exact GPS locations, size, subtype, and installation year information for rooftop solar PVs across the US. This database can be used for solar PV growth projection, solar energy integration, solar energy policy analysis and design, and spatially-resolved climate risk assessment.
- Leveraging this database, we have uncovered the socioeconomic driving factors that are correlated with earlier onset of solar adoption and higher saturated adoption levels. We have identified the heterogeneity in the effects of different types of financial incentives on solar adoption and provided implications for tailoring incentive design based on local income levels to promote equitable solar adoption.
- We have developed a distribution grid GIS mapping algorithm which can obtain granular geospatial and topology information about distribution grids using multi-modal open data, reducing the dependency on hard-to-obtain smart meter data of conventional approaches. It shows effectiveness in both the U.S. and Sub-Saharan Africa. Using this algorithm, we have uncovered the non-uniform vulnerability of distribution grids to wildfires in California in the aspects of undergrounding protection and Distributed Energy Resources (DER) preparedness. This has provided important implications for improving the affordability and equity of grid adaptation approaches.
- We have made our produced database publicly available and provided user-friendly interface to enable various stakeholders and the general public to interact with the data. We have also integrated the produced data into the Data Commons platform to enable the public to access the data and correlate it with other location-specific characteristics simply using natural language as queries.

The impact of our project is three-fold: (1) New algorithms for mapping solar PVs and distribution grids across space and time, which are open source to facilitate researchers and industry; (2) New databases of solar PVs and distribution grids that have been made publicly available for engineering, social, and policy applications; (3) New understandings and actionable insights on the potential approaches to promoting solar adoption and reducing energy infrastructure vulnerabilities.

In this report, we start by discussing the project background and motivation (section 5), followed by the overview of project objectives (section 6). Results and discussion for each task are presented in section 7. Significant accomplishments are summarized in section 8. This report will be concluded by discussing the paths forwards (section 9), products (section 10), and team roles (section 11).

4. Table of Contents

5. Background

Our energy systems are undergoing dramatic changes, including the rapid deployment of distributed energy resources (e.g., solar PVs) and increasing exposure to climate-induced extreme events (e.g., wildfires). These changes have given rise to many critical challenges: (1) Despite the introduction of various regulatory policies and financial incentives across different states in the U.S., how effective they are in promoting the adoption of distributed energy resources (DER) is largely unknown. This impedes the evidence-informed design of future policies and incentives to accelerate DER adoption in an equitable way; (2) Electric grids need to be upgraded to host a growing amount of DERs, requiring the projection of future DER adoption at a spatially-resolved level—which is challenging due to the absence of the granular spatiotemporal information of DER adoption; (3) DER-dominated electric grids can become more vulnerable to climate-induced weather extremes due to the intermittent nature of renewable DERs, but the spatial correlations among DERs, grids, and the risks of climate-induced extremes are unknown, hindering the precise investment for grid upgrades.

A fundamental gap for addressing these critical challenges is **the absence of the granular spatiotemporal information about DERs (e.g., solar PV panels) as well as their overlay with the grids and climate-induced risks**. Due to the distributed and decentralized nature of DERs, their granular spatiotemporal information is largely unavailable or dispersed in numerous “data silos” owned by different developers, utilities, or municipalities. This gap ultimately impedes the evidence-informed decision making for grid upgrades, resource allocation, and incentive design, presenting significant barriers for clean energy transition and climate change adaptation.

More specifically, despite the increasing share of solar PVs in newly added generation capacity in the U.S., we still lack an information system that maps and tracks solar adoption at a fine resolution yet large scale. This is primarily due to the decentralized nature of solar deployment. There are some previous attempts to build solar power plant databases (e.g., Global Energy Observatory [1], Global Power Plant Database [2]), but they only cover centralized solar power plants—without any information of distributed solar panels. “Tracking the Sun” database [3] used the crowdsourcing method to collect the data of distributed solar PVs in the U.S. and reported their zip-code locations. Since they rely on voluntary data contribution, they can guarantee neither completeness nor the absence of duplication.

Machine learning combined with satellite imagery offers an alternative venue for overcoming the shortcoming of the traditional solar PV data collection approaches. The availability of satellite imagery with spatial resolution less than 30 cm for the majority of the U.S., which is annually updated, offers a rich data source for solar panel detection based on machine learning. Previous pixel-wise machine learning methods [4,5] suffer from poor computational efficiency, and relatively low precision and recall (cannot reach 85% simultaneously), while previous image-wise approaches cannot provide system size information [6]. Our previous work, DeepSolar [7], used a novel deep learning approach to detect solar panels in satellite images and estimate their sizes, enabling the construction of a nationwide solar installation database for the contiguous U.S.

However, in the DeepSolar database, there is no subtype (residential, commercial, utility-scale, etc.) or temporal information (installation date) about solar panels. Moreover, this database is outdated (up to mid-2017).

In addition to the data gap for solar panels, the high-resolution information about distribution grids is also inadequate. Unlike transmission grids of which the connections and status are usually available to system operators and can be regularly measured, information about distribution grids is often incomplete, coarse-grained, or even unavailable [8]. Although utility companies may keep the information of their own distribution grids, such data are usually not publicly available or organized in a standardized format [9], OpenStreetMap maintains a spatial data collection of power lines by utilizing crowdsourcing methods, yet it is far from complete, and most of the data in this collection are for transmission lines [10].

Graph-based approaches have been previously developed for estimating distribution grid topology by leveraging measurement data from grid nodes (i.e., buses), such as smart meter measurements [11-16]. However, the applicability of these approaches is limited by the availability of smart meters, which are still not widely deployed in many places [17]. Consequently, while these graph-based techniques can identify operational topologies of grids with known measurements, they struggle with mapping complete, real-world physical grids from scratch when no prior node measurement is available. In parallel, advances in machine learning and computer vision have enabled the development of models that utilize public imagery to detect and analyze grid components. Notable efforts include using night-time light imagery to connect electrified areas and form grid maps [9], as well as using machine learning to detect poles/lines in remote sensing or street view images [18-24]. Despite these advancements, the resolution limitations of remote sensing images and the inability to map underground lines remain significant hurdles. No existing approach can construct a full distribution grid map (aboveground and underground) relying solely on publicly available data.

In this project, we aim to develop new machine-learning-based approaches to overcome the above limitations of existing methods and map solar PVs and distribution grids with ultra-high spatiotemporal granularity. This can eventually result in the construction of large-scale, high-resolution geospatial overlay of solar PVs, distribution grids, climate-induced risks, and socioeconomic attributes to enable solar incentive effect estimation, solar adoption projection, and climate risk assessment.

6. Project Objectives

This project's goal is to develop high-fidelity database of solar PVs and the infrastructure systems they rely on (i.e., distribution grids) with comprehensive and detailed information, such as temporal and subtype data. This can help bridge the critical information gap to accelerate solar adoption, to facilitate solar integration, and to mitigate climate risks of energy infrastructures, which can ultimately contribute to the national goals of clean energy transition and more sustainable and resilient economy. We achieve this by applying state-of-the-art machine learning (Convolutional Neural Networks, Siamese networks, etc.) to public and multi-modal data (e.g., satellite

imagery, street views, road networks) to obtain granular information about solar PV panels and, further, extract actionable insights. The main objectives for this project are:

1. Develop and train machine learning algorithms to identify solar panel subtype and temporal data and use it to generate detailed location, capacity, subtype (residential/commercial/utility-scale) and installation date information layers of solar installations.

2. Develop and train machine learning algorithms to identify overhead distribution infrastructure and use it to create a GIS map of the connectivity of overhead lines that excludes phase information and line parameters.

3. Form advisory group of industry specialists to examine use cases—such as planning and solar adoption forecasting - and determine a data sharing framework and best visualization methods.

4. Create web-based visualization and aggregate data sharing tool that helps navigating the produced dataset and correlating to relevant socioeconomics and policy incentives data.

The culmination of this effort is a Solar Energy Atlas that consists of a multi-layered mapping data of solar panels and distribution grid infrastructure and relevant socioeconomic, policy and irradiance information. Compared to previous labor-intensive and inefficient data collection approaches, the scalable and accurate machine-learning-based algorithms developed in this project automates and scales up the data gathering, information extraction, and knowledge discovery for solar panels, distribution grids, and their interactions with climate risks, policies, and human behaviors. This data can complement the currently used information to support utilities, vendors and analysts in applications such as substation planning and solar adoption forecasting. It also provides the granular information to support evidence-informed policy making, especially for designing incentive programs that can promote equitable solar adoption. The team works closely with an industry board of advisors to ensure the data produced can impact such applications. In addition, it can support utility, vendors, satellite data providers to establish the value of these new sources of data when limited by their resolution and availability.

This project is spread across two Budget Periods (BP). In BP1 we collect the public and partner satellite and street view data relevant to the project, use it to develop and test the machine learning models for solar panel mapping and distribution grid GIS mapping and convene a board of advisors meeting to share our progress and discuss relevant use cases. In BP2, we apply machine learning models to generate the mapping layers for selected regions in the country and develop a visualization tool for the data as well as use it to develop methodologies for industry- and policy-relevant use cases. We have also added another task (“Additional task”) beyond the original plan of Statement of Project Objectives (SOPO), which is to update the DeepSolar database to cover the solar installations up to 2023.

The summary of tasks, subtasks, as well as their planned/actual completion date is shown in Table M.1. The summary of milestones is shown in Table M.2.

Task name	Planned completion date	Actual completion date	Task summary	Actual accomplishment
Task 1: Use satellite imagery to develop and train algorithms to determine granular solar installation date	7/31/23	11/15/22	Develop machine learning (ML) models that can accurately infer the installation year of solar panels from historical satellite images. Deploy the models to California and, potentially, other states	Developed ML models that can infer installation year with 85.9% accuracy. Deployed the models to all states in the U.S.
Task 2: Solar panel subtype classification	7/31/23	11/15/22	Develop a ML model to classify each solar panel into different subtypes (residential, commercial, etc.). Deploy the models to California and, potentially, other states	Developed a ML model for subtype classification with >90% average precision and recall. Deployed the model to all states in the U.S.
Task 3: Distribution grid GIS mapping using street view imagery	7/31/23	1/15/24	Develop a ML-based model for mapping distribution grids (overhead + underground). Test/deploy the model to regions in California and countries outside the U.S.	Developed a ML-based distribution grid mapping model which can achieve >80% precision and recall in 10 regions in California. The performance can maintain a similar level when the model is transferred to countries in Africa.
Task 4: Visualization and applications development	7/31/23	4/15/23	Develop visualization platforms (both browser-based and DataCommons) for	Integrated the solar installation data into the DataCommons platform;

			the generated dataset. Use the generated dataset for solar adoption analysis.	developed a browser-based GIS platform for data visualization; Identified the non-uniform effects of incentives on solar adoption
Task 5: Data management and technical advisory board	7/31/23	3/13/24	Convene technical advisory board meetings and webinars in year 1 and year 2	Convened all four meetings; Identified new use cases and future directions (e.g., API for quick data access); Established long-term partnership with board members (e.g., Ava Community Energy)
Additional task: Updating DeepSolar dataset using moderate-to-high-resolution satellite images	7/31/23	11/15/23	Develop new ML models that can efficiently identify new solar installations and deploy the model to update the DeepSolar data to cover solar installations up to 2023	Updated the DeepSolar by incorporating solar panels installed until 2023, resulting in a dataset containing 3 million solar panels, double the amount of the old dataset

Table M.1*. Tasks, subtasks and their planned/actual completion date*

Mile-stone #	**Performance Metric**	**Targeted performance**	**Actual realized performance**	**Actual completion date**
Task 1: Use satellite imagery to develop and train algorithms to determine granular solar installation date				
1.1.1	Number of images to train the siamese network for identifying solar panel on low-resolution (LR) images	≥10,000	A dataset with 56,429 images for model training	10/31/21

Mile-stone #	Performance Metric	Targeted performance	Actual realized performance	Actual completion date
1.2.1	The precision and recall of solar panel detection in low-resolution (LR) images	>90%	Recall: 91.2% Specificity: 95.6%	10/31/21
1.2.2	The correctness rate of solar panel Installation year inference	>80%	The correctness rate is 85.9±1.0%	1/31/22
1.3.1	The state(s) with full coverage of temporal information (installation year) of solar panels	California	Have obtained the temporal information for all 50 states and the D.C. in the U.S.	10/31/22
Task 2: Solar panel subtype classification				
2.1.1	Number of images to train the solar subtype classifier	≥10,000	A dataset with 12,948 images for model training	10/31/21
2.2.1	The micro- and macro-average of precision and recall of subtype classification	>0.75	Macro-average: Precision: 0.908 Recall: 0.917 Micro-average: Precision: 0.917 Recall: 0.917	1/31/22
2.3.1	The state(s) with full coverage of subtype information of solar panels	California	Have obtained the temporal information for all 50 states and the	7/31/22

Mile-stone #	Performance Metric	Targeted performance	Actual realized performance	Actual completion date
			D.C. in the U.S.	
Task 3: Distribution grid GIS mapping using street view imagery				
3.1.1	Number of images to train the models for identifying power lines and poles	≥ 8000	A dataset with 10,000 images for model training	10/31/21
3.1.2	Precision and recall for both line and pole detection	>0.85	Line detection: Precision: 0.982 Recall: 0.937 Pole detection: Precision: 0.982 Recall: 0.851	10/31/21
3.2.1	Pole localization accuracy (in ratio of actual poles that can be detected within 25m)	>80%	Pole localization precision: 83.2% recall: 83.6%	1/31/22
3.2.2	Precision and recall in link prediction	>70%	Link prediction precision: 78.7% Recall: 76.6%.	1/31/22
3.3.1	Precision and recall for the similarity comparison between actual PG&E distribution grid map (above- plus underground)	>80%	Compared with PG&E grid, the model achieved recall: 83%	4/30/22

Mile-stone #	Performance Metric	Targeted performance	Actual realized performance	Actual completion date
			Precision: 89%	
3.3.2	Number of missing lines in the utility-owned grid maps that can be identified and corrected with the grid mapping model.	>0	Identified 9 to 132 missing lines (across different test areas).	1/31/22
3.4.1	Number of regions in California to deploy the grid GIS mapping model.	≥10	The model has been deployed to 10 regions in California	1/31/23
3.4.2	The decrease in precision and recall of network link prediction when transferring the model to another region outside US.	<15%	The decrease in precision and recall is less than 7% when transferring the model to other regions	7/31/22
3.5.1	The precision and recall of the distribution grid GIS mapping algorithm, compared with detailed GIS maps provided by utility partners	>80%	GIS mapping recall: 85% Precision: 90%	3/31/24
Task 4: Visualization and applications development				
4.1.1	Number of types of data to integrate and display at aggregate level on the browser-based platform.	≥ 5	11 types of data layers have been integrated and displayed on the platform	4/30/22
4.2.1	Obtain a roadmap for the development of the schema for	Complete = TRUE	Complete = TRUE	4/30/22

Mile-stone #	Performance Metric	Targeted performance	Actual realized performance	Actual completion date
	Energy Data Commons with a focus on Solar Energy Atlas			
4.3.1	Regression R^2 on out-of-sample test set for solar adoption prediction	>0.5	Achieved a R^2 of 0.65 for solar adoption prediction	4/30/23
4.4.1	The coverage of overhead line ratio and solar PV capacity for very-high-fire-risk regions in PG&E territory in California.	100%	Covered 100% very-high-fire-risk regions in PG&E territory with overhead line ratio and solar PV capacity information	10/31/22
4.5.1	Number of mismatches within numeric tolerance between Solar Energy Atlas on DataCommons and its from original offline version.	0	0 mismatch between online and offline versions	4/30/23
4.6.1	Upload Solar Energy Atlas data to Data Commons and test integration	Complete = TRUE	Complete = TRUE	4/30/23
Task 5: Data management and technical advisory board				
5.1.1	Convene industry advisory board meeting in Year 1	Complete = TRUE	Complete = TRUE	4/22/22
5.2.1	Convene a webinar in Year 1 on computer vision applications for grid	Complete = TRUE	Complete = TRUE	8/15/22
5.3.1	Convene industry advisory board meeting in Year 2	Complete = TRUE	Complete = TRUE	4/30/23

Mile-stone #	Performance Metric	Targeted performance	Actual realized performance	Actual completion date
5.4.1	Convene a webinar in Year 2 on discuss applications using the Solar Energy Atlas	Complete = TRUE	Complete = TRUE	3/13/24
Go/No-Go decision point				
G/NG 1A	Percentage of milestones achieved in Year 1	100%	Achieved 100% of the milestones in Year 1	7/31/22
G/NG 1B	Convene both industry advisory board meeting and webinar for Year 1	Complete = TRUE	Complete = TRUE	7/31/22

Table M.2*. Milestones and Go/No-Go decision points*

7. Project Results and Discussion

This section quantitatively presents the project results and discussion. It is organized by tasks (Task 1 to 5, as well as the Additional Task) and subtasks. For each task/subtask, we start by introducing its overall goal, followed by the technical discussion of every milestone (anticipated outcomes vs. realized outcomes) in each subtask.

7.1. Task 1: Develop and train algorithm to determine granular solar installation date using historical satellite imagery

The goal of this task is to develop the algorithms for determining the installation year for solar PVs using historical satellite images. This is used for constructing a nationwide solar PV installation database with granular spatial (GPS location) and temporal information. The major challenge here is the low image resolution of historical satellite images, which is tackled in Subtask 1.1 with the development of a pseudo-siamese neural network. This model is benchmarked against the manually-curated test set (Subtask 1.2). After the model development and extensive evaluation, the model is deployed to construct a nationwide spatiotemporal solar PV installation database (Subtask 1.3).

7.1.1. Subtask 1.1*:* Solar panel identification in low-resolution historical satellite imagery

Due to the low-resolution of some historical satellite images, directly applying the original DeepSolar model [7] that was trained using high-resolution images to the historical images can yield unsatisfactory results. Therefore, the goal of this subtask is to overcome the low-resolution challenge of historical satellite imagery by developing

novel machine learning models. To facilitate the model training and testing, a large-scale dataset with manually-verified labels is needed. Below we introduce the dataset construction and machine learning model design.

	Positive	Negative
Training	11623	39175
Validation	481	1562
Test	806	2782

Table T1.1: Number of samples in training/validation/test sets for low-resolution (LR) solar system identification dataset. *Positive: contain solar. Negative: no solar.*

Milestone 1.1.1: Collect ≥10,000 images to train a Siamese neural network for identifying solar panels in low-resolution images

We have achieved this milestone Q1 of BP1. Specifically, we have constructed **a dataset with 56,429 images in total (target number: 10,000 images)** to train deep learning models for identifying solar panels in low-resolution historical satellite images. The dataset is partitioned into training, validation, and test sets. The images were retrieved using Google Earth and they are divided into two classes by manual checking: contain solar panel (positive) and no solar panel (negative). Table T1.1 shows more details about the different partitions of this dataset.

To facilitate solar identification in low-resolution (LR) images, we developed a two-branch pseudo-Siamese Convolutional Neural Network (CNN) that takes a target LR image and a “reference” positive high-resolution (HR) image as inputs, and outputs the score indicating whether the target LR image contains solar. Inspired by the visual tracking models in the computer vision field [24], we develop a two-branch CNN with each branch taking either the target LR image or its reference HR image as inputs. The two branches have identical architecture but different weights hence the model is called “pseudo-Siamese”. By comparing the feature maps generated by each of the two branches, the model is able to estimate the similarity between the LR target image and its HR reference image. To tackle the potential object displacement between target and reference image, a cross-correlation operator is utilized to compare the feature maps from each of the two branches. Based on the similarity, the model finally outputs the score indicating the probability of the target LR image containing solar. In this work, we use ResNet-34 network [25] as the backbone for each branch. The overall model architecture is shown in Figure 1.1. Feature maps after the 2nd, 3rd, and 4th stack of building blocks in the ResNet-34 are used for comparison. Depth-wise cross-correlation operator, which compares the features from different channels separately, is applied to each of the three pairs of feature maps and generates a similarity map for each. All three similarity maps are concatenated together and then fed into three convolutional layers in a series and finally output the logit score.

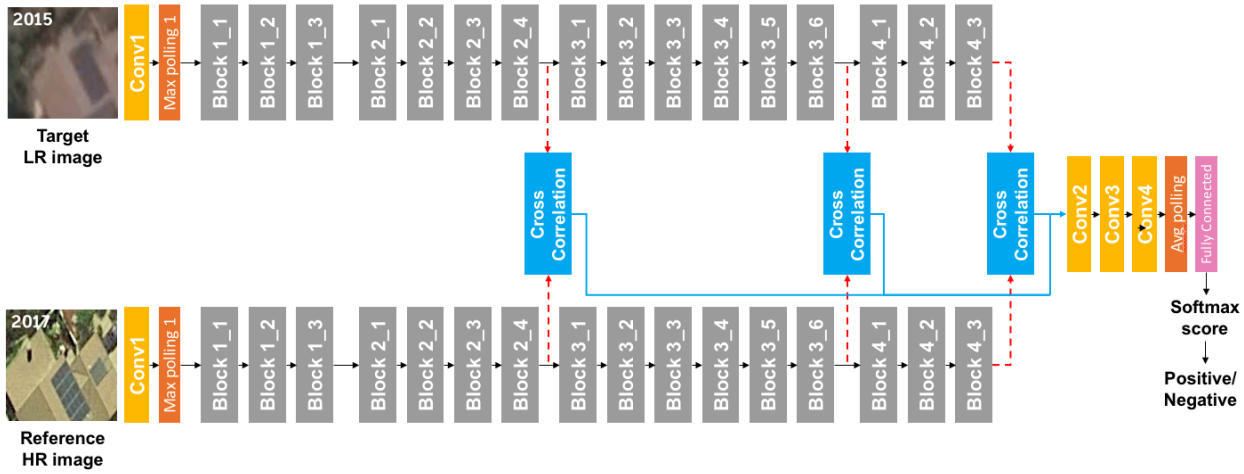

Figure 1.1: The pseudo-siamese network for solar identification in low-resolution (LR) images. *Each branch is a ResNet-34 network which takes either target LR image or reference HR image as inputs. Feature maps after the 2nd, 3rd, and 4th stacks of building blocks are compared using cross-correlation modules and then concatenated. ReLU and batch normalization layers are not shown in this figure.*

7.1.2. Subtask 1.2: Benchmark the performance of the low-resolution solar identification methodology

The goal of this subtask is to evaluate the performance of the machine learning model for low-resolution solar PV identification as well as the entire pipeline for determining the solar PV installation year. The evaluation is performed at the two levels: the image level (Milestone 1.2.1) and the system level (Milestone 1.2.2). Below we detail the evaluation method and metric values for each of them.

Milestone 1.2.1: Achieve a low-resolution (LR) solar panel detection precision and recall ≥ 90%

We have achieved this milestone in Q1 of BP1. We test the performance of solar panel detection on the test set partition (3,588 images) constructed in Milestone 1.1.1. The model achieves a sensitivity/recall (true positive rate) of **91.2%** and a specificity of **95.6%** (true negative rate) on this test set. Both of them are higher than the target value **90%.**

Explanation of variance: we use sensitivity (another name of recall) and specificity as the metrics instead of recall and precision (proposed in the original Milestone 1.2.1 in SOPO). This is because sensitivity and specificity are directly related to our final target metrics, the correctness rate of predicting year of PV installations.

Specifically, if we ignore the rare cases that solar panel can be uninstalled later after installation, and use “0” to denote negative sample and “1” to denote positive sample, our image sequence is a sequence with all “0” in the first part and all “1” in the last part. There cannot be “1” between “0” such as “00100”. For a single image whose ground-truth label is 0, its probability of being predicted correctly is:

$$P_0 = TN/(TN+FP) = \text{specificity}$$

And for a single image whose ground-truth label is 1, its probability of being predicted correctly is:

$$P_1 = TP/(TP+FN) = \text{sensitivity}$$

If we have an image sequence with first *n* images as negative and last *m* images as positive, then the probability of predicting the whole sequence correctly is:

$$P_{sequence} = P_0^n P_1^m = (\text{specificity})^n\ (\text{sensitivity})^m$$

Therefore, optimizing specificity and sensitivity at the image level can directly improve the correct rate of predicting installation year at the sequence level, hence we use sensitivity/recall and specificity as targeted metrics at the image level instead of recall and precision.

	HR	LR	Extremely blurred
Training	11844	17178	1585
Validation	2010	2043	182
Test	4340	3588	553

Table T1.2: Number of samples in training/validation/test sets for blur detection dataset. HR: high resolution. *LR: low resolution.*

	Positive	Negative
Training	7148	4696
Validation	1189	821
Test	2392	1948

Table T1.3: Number of samples in training/validation/test sets for high-resolution (HR) solar system identification dataset. *Positive: contain solar. Negative: no solar.*

Milestone 1.2.2: At the system level, achieve a correctness rate ≥ 80% for solar installation year prediction

We have achieved this milestone in Q2 of BP1. We deploy the well-trained models on a out-of-sample sequence test set containing 1,164 image sequences and compare the predicted year of installation with actual year of installation. The correctness rate of installation year prediction (ratio of sequences with predicted installation year equal to the actual installation year) is **85.9±1.0% (target value: 80%)**. Below we elaborate on (1) the datasets used for model development and testing (in addition to the LR dataset

introduced in Milestone 1.1.1) and (2) the overall framework of installation year prediction.

Datasets: For a solar system recorded in the DeepSolar database, we retrieve a sequence of images at its geolocation with each image captured in a year between 2005 and 2017. In addition to the low-resolution (LR) image dataset introduced in Milestone 1.1.1, we construct two other image datasets with image-wise labels to develop different modules of the framework, as well as an image sequence test set for evaluating the overall accuracy of predicting year of installation. Specifically, to develop the blur detection model to determine the resolution of an image, we construct an image dataset containing 43,323 images with three classes—high resolution (HR), low resolution (LR), and extremely blurred (See their detailed statistics in Table T1.2); to develop the solar system identification model for HR images, we construct an image dataset containing 18,194 HR images with binary labels indicating whether a solar system exists in an image (See their detailed statistics in Table T1.3). Samples in these image datasets are randomly selected across 11 counties from 9 states. Moreover, the image sequence dataset contains 238 sequences for validation and 1,164 sequences for testing. Each sequence is manually labeled with the installation year of the PV system by visual inspection as its ground truth. Besides the 11 counties included in the image datasets, the image sequence test set covers additional 12 counties from 10 states. Samples in training, validation, and test set are mutually exclusive.

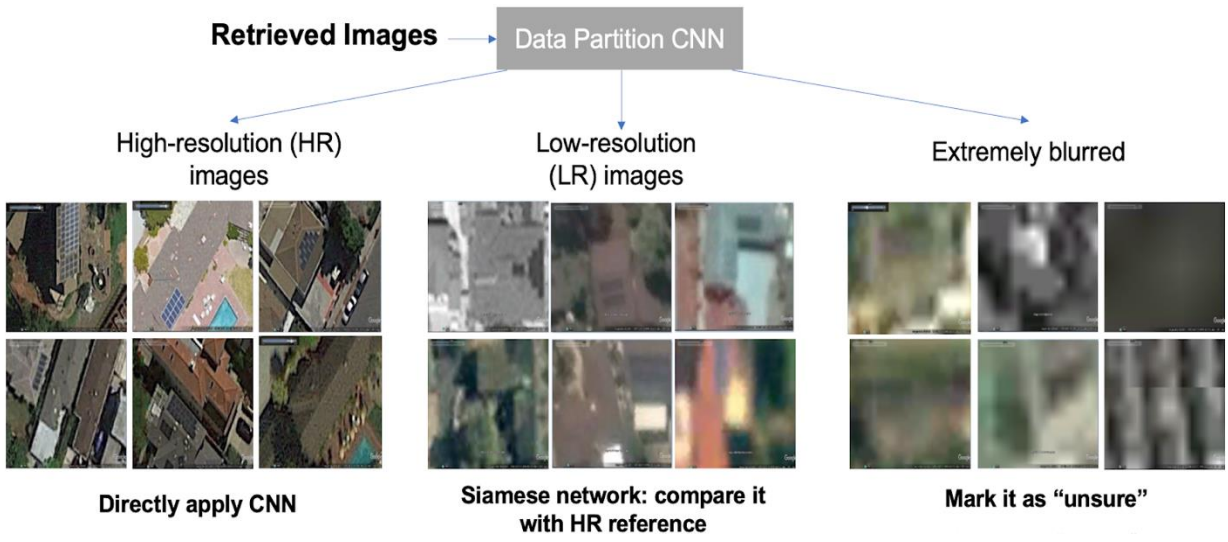

Figure 1.2: Overall framework that first classifies images according to their resolution into three classes: high-resolution (HR), low-resolution (LR), and extremely blurred. *We use a normal single-branch Convolutional Neural Network (CNN) to process the HR images, and use the siamese network (two-branch CNN) we developed to process the LR images. For extremely blurred images, we mark them as uncertain so that it will not be used for predicting year of installation, since it is out of the distribution of the HR or LR image training set.*

Overall framework for determining installation year: The overall framework of image processing is shown in Figure 1.2. Before identifying solar systems in a satellite image,

we first leverage a Convolutional Neural Network (CNN) to classify the image into one of the three classes according to its resolution — HR, LR, and extremely blurred.

For an extremely blurred image, we mark it as uncertain so that it will not be used for predicting year of installation, since it is out of the distribution of the HR or LR image training set. For HR images, we directly apply an Inception-v3 network on HR images and it outputs a score indicating the probability of an input image containing PV. The model can achieve a sensitivity (true positive rate) of 97.6% and a specificity (true negative rate) of 98.5% on the test set. For LR images, we apply the two-branch pseudo-Siamese CNN which has been introduced in Milestone 1.2.1.

In deployment, given a sequence of historical satellite images for a PV system, we run the HR model on all images and use all positively classified images as the reference images. Whether an HR image contains PV is determined by the classification result generated by the HR model. A LR image is predicted as containing PV if any pair of a reference image and itself gets a positive prediction by the LR model. The first year when positive images appear is predicted as the year of installation of the PV system.

7.1.3. Subtask 1.3: Run the solar identification model on solar installation records in the DeepSolar database

The goal of this subtask is to deploy the model we developed to determine the year of installation for every solar PV documented in the DeepSolar database. The eventual outcome is a nationwide solar PV installation database with temporal information. Below we introduce our accomplishment.

Milestone 1.3.1: Use the newly-developed models to obtain the installation year information for solar PVs in California

We have achieved this milestone in Q1 of BP2. We have downloaded image sequences for all residential and commercial solar PV installations not just in California, but across the U.S. (1,057,070 systems). We have applied the installation year prediction model developed in Subtask 1 and 2 to each of these image sequences and obtained their installation year information. **Note that our initial target of the data coverage (as proposed in the original Milestone 1.3.1) is California, while our actual realized coverage is the U.S.** The overall temporal variation of solar adoption rate, characterized by the number of solar installations per 1000 households, is shown in Figure 1.3.

7.2. Task 2: Solar panel subtype classification

The granular information about solar installation subtypes is absent in the existing solar installation database (i.e., DeepSolar). In this task, we aim to develop a machine learning model to identify the subtype for solar installations from satellite imagery and fill out the blank of such information in the database.

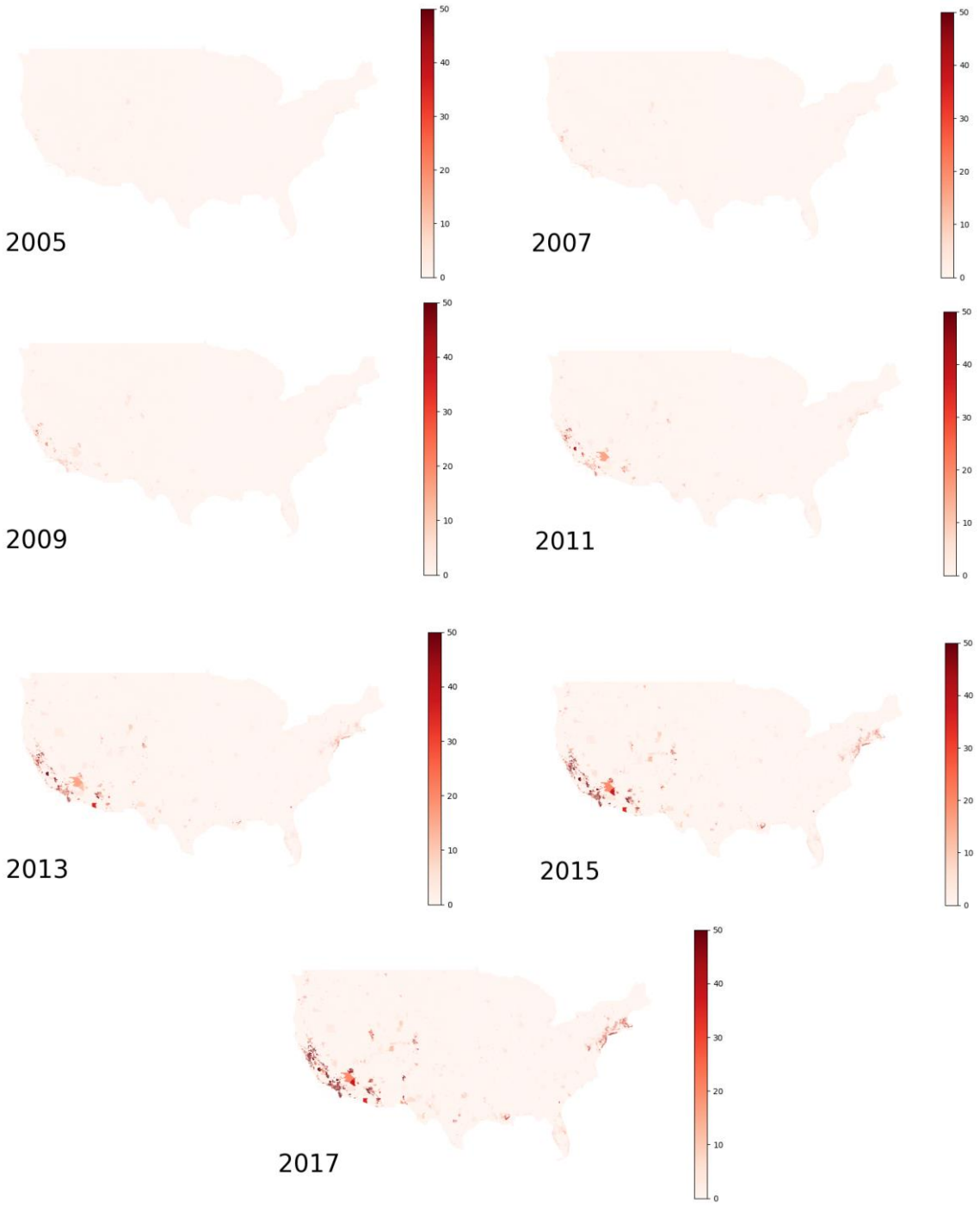

Figure 1.3: Temporal variation of residential solar adoption rate at the census block group level across the contiguous US. *The residential solar adoption rate is characterized by the cumulative number of residential solar PVs per 1000 households in a census block group.*

7.2.1. Subtask 2.1: Solar panel subtype classification

The goal of this subtask is to develop a machine learning model to automatically classify each solar panel into different subtypes, including residential, commercial, utility-scale, and solar water heating. Some non-solar-panel objects were wrongly identified by the

original DeepSolar model as solar panels, so they also need to be filtered out. We approach this goal by developing both the dataset for training and testing as well as the Convolutional Neural Network (CNN) model for ordinal classification.

Milestone 2.1.1: Collect ≥ 10,000 images to train a solar subtype classifier
We have achieved this milestone in Q1 of BP1. We have constructed a dataset of **12,948 images (target number: 10,000 images)** in total with solar subtype labels (utility-scale PV, commercial PV, residential PV, solar heating, and negative samples). Note that we included a "negative" subtype, as the original DeepSolar produced a small fraction of false positive samples (which are actually negative samples) which need to be filtered out. The dataset is partitioned into training/validation/test sets. These images are randomly sampled from the solar installation records in DeepSolar dataset. The details about the dataset are shown in Table T2.1. Below, we further introduce the solar subtype classifier.

	Training set	Validation set	Test set
Utility-scale PV	404	72	73
Commercial PV	1301	194	194
Residential PV	4399	588	589
Solar heating	1559	298	298
Negative	2338	320	321

Table 2.1: Number of samples in training/validation/test sets for solar panel subtype classification. *Training, validation, and test set are mutually exclusive.*

Figure 2.1: visualization of samples in the dataset for solar panel subtype classification.

The characteristics of solar subtypes are ordinal. Specifically, utility-scale PVs usually have relatively large sizes; commercial PVs are usually smaller than utility-scale PVs, but usually larger than residential PVs; residential PVs and solar heating systems are installed on rooftops of residential buildings; utility-scale, commercial, and residential PVs are photovoltaic systems for generating electricity while solar heating is not. Therefore, we can order these five subtypes to facilitate the training of a CNN model (ResNet-50): utility-scale PV -> commercial PV -> residential PV -> solar heating -> negative. We can assign ordinal multi-class labels to enforce such ordinal relationships during training:

- Utility-scale PV: [1, 1, 1, 1]
- Commercial PV: [1, 1, 1, 0]
- Residential PV: [1, 1, 0, 0]
- Solar heating system: [1, 0, 0, 0]
- Negative: [0, 0, 0, 0]

In this way, the penalty of misclassifying a residential PV into utility-scale PV is higher than misclassifying it into commercial PV. Such an ordinal relationship can provide extra guidance for models to extract useful information from visual features for subtype classification.

In practice, the subtype can be determined based on the prediction score of an image $[x_1, x_2, x_3, x_4]$:

- If $x_1 < \text{threshold}_1$: “negative” (falsely detected by the original DeepSolar model)
- Else if $x_2 < \text{threshold}_2$: “solar water heating system”
- Else if $x_3 < \text{threshold}_3$: “residential PV”
- Else if $x_4 < \text{threshold}_4$: “commercial PV”
- Else: “utility-scale PV”

The thresholds are determined by the performance on the validation set.

Subtask 2.2: Benchmark the performance of the solar panel subtype classifier
The goal of this subtask is to evaluate the performance of the CNN model for solar panel subtype classification on an out-of-sample test set with ≥1000 images. These images are not used for training.

Milestone 2.2.1: Achieve micro- and macro-average of precision and recall of subtype classification ≥ 0.75
We have achieved this milestone in Q2 of BP1. Specifically, we run the well-trained solar subtype classification model (developed in Subtask 2.1) on the out-of-sample test set with 1,475 images (constructed in Subtask 2.1) and compare the model outputs with the ground-truth labels. The confusion matrix of the model performance is shown in Figure 2.2. We further calculate the macro- and micro-average of the precision and recall for evaluating the overall performance of the multi-class classification. The definition of macro-average and micro-average precision and recall are:

$$Macro\ avg.precision = \frac{1}{N}\sum_{i=1}^{N} precision_i$$

$$Micro\ avg.precision = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} TP_i + \sum_{i=1}^{N} FP_i}$$

$$Macro\ avg.recall = \frac{1}{N}\sum_{i=1}^{N} recall_i$$

$$Micro\ avg.recall = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} TP_i + \sum_{i=1}^{N} FN_i}$$

Here, N is the number of subtypes (5 in our case). TP_i, FP_i, and FN_i are the numbers of true positive, false positive, and false negative samples of subtype i, respectively. $precision_i = TP_i / (TP_i + FP_i)$. $recall_i = TP_i / (TP_i + FN_i)$.

Based on the confusion matrix, we can calculate the macro-average of precision and recall, which are **0.908** and **0.917**, respectively. We also calculate the micro-average of precision and recall, which are **0.917** and **0.917**, respectively. They are all above the target value **0.75**.

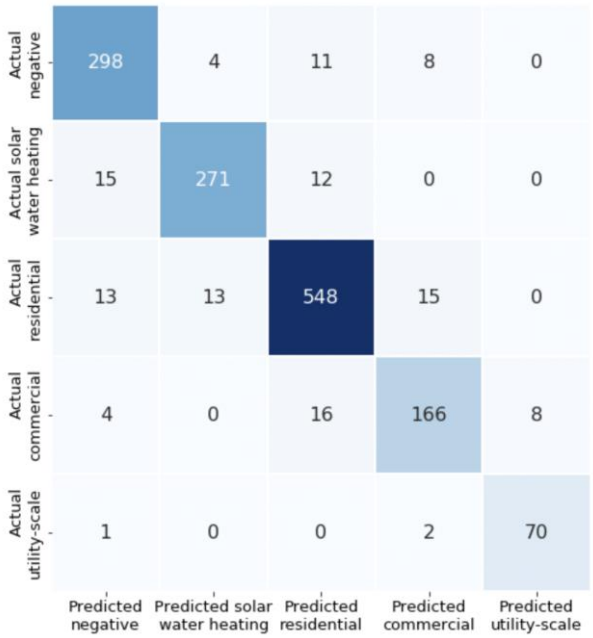

Figure 2.2: Confusion matrix of the solar panel subtype classification on the test set. *The number in each cell is the number of samples in each prediction category. For example, the number at the 2nd row and 1st column is 15, which means 15 of the actual solar water heating systems are predicted to be negative.*

7.2.3. Subtask 2.3: Run the solar panel subtype classifier model on all solar installations in the DeepSolar database

The goal of this subtask is to deploy the solar panel subtype classification model we developed for every solar PV documented in the DeepSolar database. The eventual outcome is a nationwide solar PV installation database with solar panel subtype information. Below we introduce our accomplishment for milestone 2.3.1.

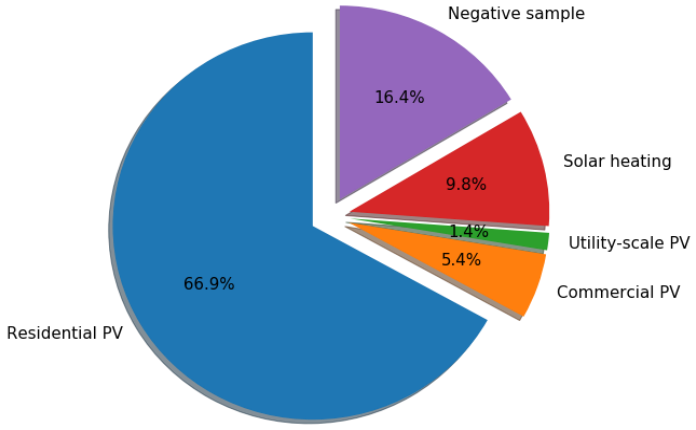

Figure 2.3: the fraction of each subtype of solar installations recorded in DeepSolar dataset, obtained by the solar subtype classifier. *Negative samples are those falsely identified as positive by the original DeepSolar model but corrected by the solar subtype classifier.*

Milestone 2.3.1: Obtain subtype information for 100% of solar installation records in California in DeepSolar database

We have achieved this milestone in Q4 of BP1. Specifically, we apply the solar panel subtype classification model to **all solar installations recorded in the DeepSolar database**. The model is applied on the latest remote sensing high-resolution (<10cm) captured at the geolocation of each recorded solar installation. We finally obtain the subtype information for each of these solar installations. Among all 1,470,189 records in the original DeepSolar dataset, there are 983,970 residential PVs, 80,088 commercial PVs, 21,183 utility-scale PVs, and 144,147 solar water heating systems. The remaining 240,801 systems are negative samples which are falsely identified as positive by the original DeepSolar model. Figure 2.3 shows the fraction of each subtype (including the negative samples) determined by the solar subtype classifier. **Note that our initial target of the data coverage (as proposed in the original Milestone 2.3.1) is California, while our actual realized coverage is the U.S.**

7.3. Task 3: Distribution grid GIS mapping using street view imagery

In this task, we aim to develop a distribution network GIS mapping tool with machine learning. The goal is to develop machine learning models that can detect both utility

poles and power lines in street view images, and combine them with publicly-available road network and building data to estimate the distribution grid GIS map.

7.3.1. Subtask 3.1: Power line detection and utility pole detection

The goal of this subtask is to develop machine learning models to detect power lines and utility poles from street view images. This includes street view image dataset construction (Milestone 3.1.1), the development of power line detector and utility pole locator, and model performance evaluation (Milestone 3.1.2).

Milestone 3.1.1: Collect ≥ 8,000 images to train the models for identifying power lines and poles

We have achieved this milestone in Q1 of BP1. We have constructed a street view image dataset to train the line detector and pole detector. The dataset contains **10,000 upward satellite images (target number: 8,000 images)** which are randomly sampled from the San Francisco Bay Area. Each image is annotated with two labels indicating whether it contains lines and whether it contains poles respectively. There are 3,204 images containing line(s) and 1,786 images containing pole(s). The dataset is split into training, validation, and test sets following the 85%-7.5%-7.5% ratio.

Milestone 3.1.2: Achieve precision and recall for both line and pole detection > 0.85

We have achieved this milestone in Q1 of BP1. For line detection, the model achieves a precision of **0.982** and a recall of **0.937** on the test set. For pole detection, the model achieves a precision of **0.982** and a recall of **0.851** on the test set. **They are all above the target value 0.85.** Below we elaborate on the power line detection and utility pole detection models.

Each upward street view image is processed by two CNNs—a power line detector and a utility pole detector. The line detector classifies an image into positive (contain lines) or negative category (no line found), and then extracts the line directions for positive images (Figure 3.1A). Similarly, the pole detector classifies the image first and then estimates the pole orientations (Figure 3.1B). Both models adopt an Inception-v3 model architecture (Figure 3.2).

To estimate the directions of power lines in an image, we apply Hough transform on Class Activation Maps (CAMs) generated by the line detector. Hough transform can detect a line and estimate its direction in a CAM. In order to tackle multiple lines in an image, once a line is detected, we hide it by adding a mask to the CAM and re-apply Hough transform to it, until all lines in the image have been detected. Similarly, for estimating pole orientations, we also apply Hough transform on the CAM generated by the pole detector and calculate the angle between the pole and horizontal axis of the image.

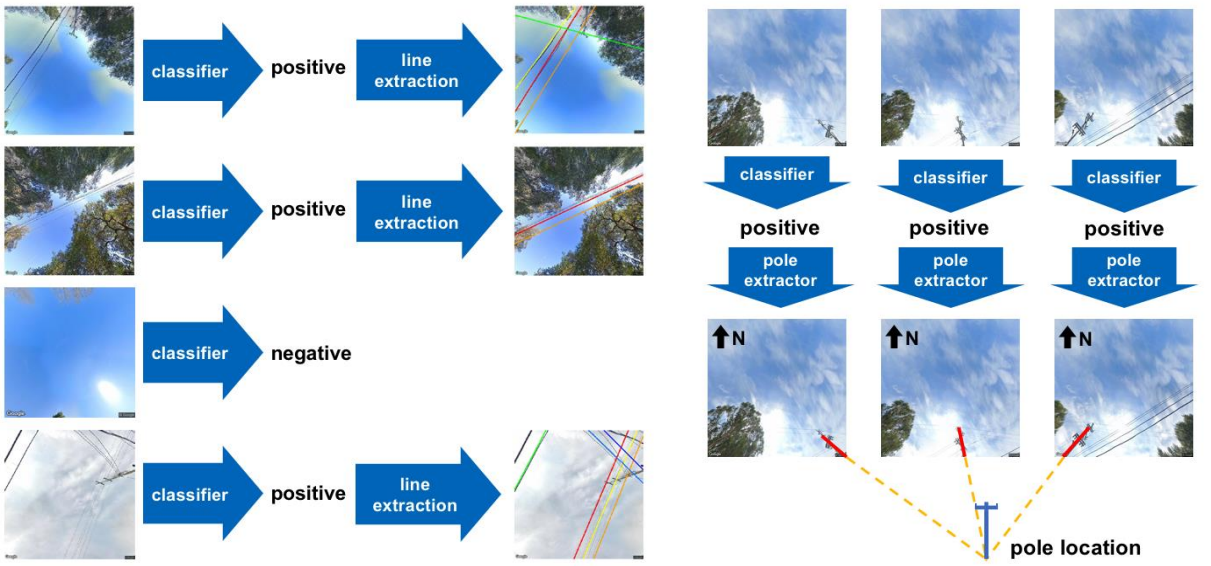

Figure 3.1: Power line detector and utility pole locator. *(A) Power line detector. It first decides whether there is any line in the image (positive) or not (negative), and then extracts lines in positive images and estimates their directions using Hough Transform. (B) Utility pole detector. It first decides whether there is any pole in the image (positive) or not (negative), estimates the pole orientations, and then intersects rays of pole orientations to obtain the pole location.*

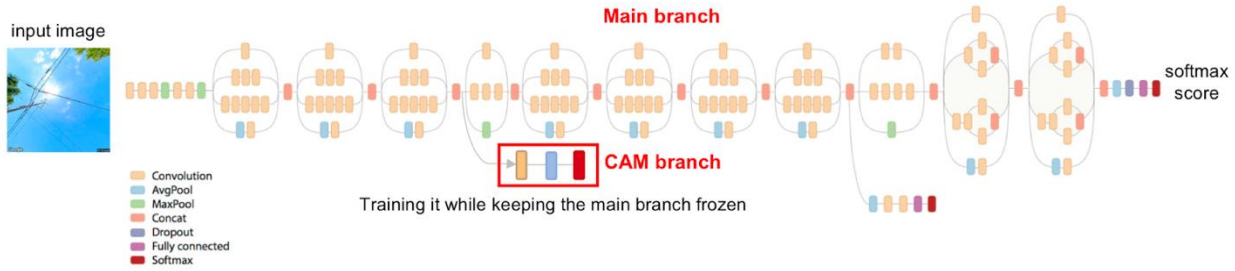

Figure 3.2: The model architecture for both line detector and pole detector. *A segmentation branch is added at an intermediate layer of the Inception-v3 network to generate Class Activation Map (CAM).*

To estimate the exact geo-coordinates of poles, we assume utility poles are approximately vertical, hence any poles in an upward street view must point to the image center. Under this assumption, by drawing rays of pole orientations starting from street view points and intersecting them, the exact locations of poles can be derived (Figure 3.1(B)). Intersecting two rays can obtain a single intersection point, while intersecting three or more rays can potentially obtain multiple intersections and we use spatial clustering to merge intersections that are close to each other.

7.3.2. Subtask 3.2: GIS mapping using link prediction

The goal of this subtask is to combine the detected power lines and utility poles (achieved by the machine learning models developed in Subtask 3.1) with the road network information to build a GIS mapping pipeline for distribution grids. The essential part of this subtask is the development and evaluation of a machine-learning-based link prediction model, which is detailed as follows. Milestone 3.2.1 and Milestone 3.2.2 are designed for evaluating the pole localization and link prediction performance, respectively.

Link prediction method: By incorporating road information, we aim to further improve the prediction of whether there is a line connection between two detected poles. Specifically, each road instance can be represented as a series of line segments. If a detected pole or a street view point has a distance ≤ D_{attach} to a road, it will be attached to that road. All attached street view points and detected poles are sorted in order along the road. In this way, features for a pair of poles, such as whether the two poles are next to each other along the road, whether they are attached to the same road, whether there are street view points with power lines detected between them, etc, can be extracted from the road model.

Moreover, to reduce the number of poles missed by the pole detector, we insert pole(s) between a pair of poles if the distance between them is greater than a threshold D_{insert}.

Road maps, which can be represented as geospatial graphs with nodes and edges, are obtained from OpenStreetMap [10].

We develop a machine-learning-based link prediction model that takes feature variables for a pair of poles as inputs and outputs whether there is a line connection between them. Any pair of poles with distance less than a threshold D_{cand} are considered as candidates. We consider various types of classification models including logistic regression, decision tree, random forest, support vector machine, and gradient boosting. Feature candidates include:

- Distance between the two poles.
- Whether the two poles are on the same road.
- Whether the two poles are next to each other along the road.
- Ratio of street view points with power line detected between the two poles.
- Minimum/average difference between the line directions estimated from street view images and the direction of the line connecting the two poles. Small difference gives evidence that there are power lines between the two poles.
- Whether either of the poles is detected by the pole detector or inserted.
- Whether either of the poles is at a road intersection.
- Whether the two poles are at the same road intersection.
- The binary prediction of a modified Dijkstra's algorithm [9]. This algorithm finds the most efficient paths to connect poles. On the meshed spatial map, each cell is assigned with a weight. By setting the weights of roads to be lower than others, connecting poles along the road is preferrable.

We use cross-validation on the development set to select the best model as well as the best feature sets. The output of the link prediction module is a geospatial graph with estimated geotagged poles as nodes and predicted line connections as edges.

Dataset: To develop the link prediction model and evaluate the overall grid mapping performance, we collect and clean distribution grid maps in 6 different regions and treat them as ground truth maps. The 6 regions are from cities in Northern California including San Carlos, Newark, Santa Cruz, Yuba City, Pacific Grove, and Salinas. For these 6 regions, we obtain the geospatial maps of distribution grids from the Integration Capacity Analysis (ICA) map [26] of Pacific Gas and Electric Company (PG&E), and then manually distinguish between overhead and underground grids. For the geospatial graph of overhead grids, we only keep nodes that are corresponding to utility poles and edges that are corresponding to power lines by checking other data sources such as satellite images and street view images. Grid map in Santa Carlos is used as a development set for training and validating the link prediction model while the grid maps in other 5 regions in Northern California are used as test sets.

Evaluation metrics: We evaluate the performance of pole localization and link prediction. To compare a set of ground truth geolocations of poles $P = \{p_1, p_2, \dots p_M\}$ and a set of estimated geolocations of poles $Q = \{q_1, q_2, \dots q_N\}$, we match all pairs of poles from two sets $\{(p_i, q_j\}\}_{1 \le i \le M,\ 1 \le j \le N}$ and sort them in ascending order according to the distance between the pair of poles. Given a distance threshold $D_{matching}$, we pick pairs out of the sorted list starting from the first element and add them to the list of matched pairs until the pairwise distance becomes greater than $D_{matching}$. If either estimated or ground truth pole in a pair have already been picked before, this pair will be dropped and not picked again to avoid repetition. Then we use the precision and recall for measuring the pole localization performance, defined as:

$$precision\ of\ pole\ localization = \frac{\#\ matched\ pairs}{N}$$

$$\text{recall of pole localization} = \frac{\text{\# matched pairs}}{\text{M}}$$

To evaluate the link prediction performance of overhead grids, we compare the ground truth edge set E and the edge set generated by the link prediction model F. Specifically, we define the precision and recall for link prediction as:

$$\text{precision of link prediction} = \frac{|\text{E} \cap \text{F}|}{|\text{F}|}$$

$$\text{recall of link prediction} = \frac{|\text{E} \cap \text{F}|}{|\text{E}|}$$

Here || denotes the number of edges in a set. ∩ denotes the intersection of two sets.

Note that edges between false negative poles (poles that are not detected) are counted as false negative edges, and edges between false positive poles (wrongly-detected poles) are counted as false positive edges. Moreover, false negative or false positive poles between two true positive poles along the same power lines do not affect the overall topology. For example, q_m–q_n in the predicted edge set F can be viewed as a correct prediction for p_i–p_j–p_k in the true edge set E if p_i matches q_m and p_k matches q_n. q_l–q_m–q_n in F can also be viewed as a correct link prediction for p_i–p_j in E if p_i matches q_l and p_j matches q_n. To give tolerance to such errors, we measure the precision and recall after matching the equivalent segments from E and F.

The parameter values used in our experiment are D_{attach} = 20m, D_{insert} = 70m, and D_{cand} = 100m. We select the best feature set and parameters for link prediction models based on the 9-fold cross-validation on the San Carlos development set which are divided into 9 subsets according to the boundary division of 9 census tracts. The model with the best configuration is then trained on the full development set and applied on all test areas.

Milestone 3.2.1: Achieve pole localization accuracy (in ratio of actual poles that can be detected within 25m) > 80%

We have achieved this milestone in Q2 of BP1. The pole localization method is introduced at the end of subsection 7.3.1. The evaluation metrics are detailed in subsection 7.3.2. Table T3.1 (column 1 and column 2) shows the pole localization performance with $D_{matching}$ = 25m. Compared with the ground truth pole locations derived from the PG&E ICA map, for most of the test areas, over 80% of the actual poles can be detected within 25m (recall) while over 80% of the detected poles have a nearby actual pole within 25m (precision). The average precision and recall over all 5 test areas (excluding the development set of Santa Carlos) are **0.832** and **0.836**, respectively. They are both higher than the **target value 0.8 (80%)**.

Milestone 3.2.2: Achieve precision and recall pf link prediction > 70%

We have achieved this milestone in Q2 of BP1. The link prediction method and the evaluation metrics are introduced in subsection 7.3.2. We compare the performances of two different link prediction models—decision tree and gradient boosting—on the test areas in Northern California (introduced in subsection 7.3.2), and the result shows that gradient boosting performs slightly better than decision tree in terms of F1 score (see Figure 3.3). For the gradient boosting model (see Table T3.2), the precision after matching equivalent segments ranges from 0.71 to 0.83 in the 5 test areas, while the recall ranges from 0.67 to 0.89 (Table T3). The average precision and recall over all 5 test areas (excluding the development set of Santa Carlos) are **0.787** and **0.766**, respectively. They are both higher than the **target value 0.7 (70%)**.

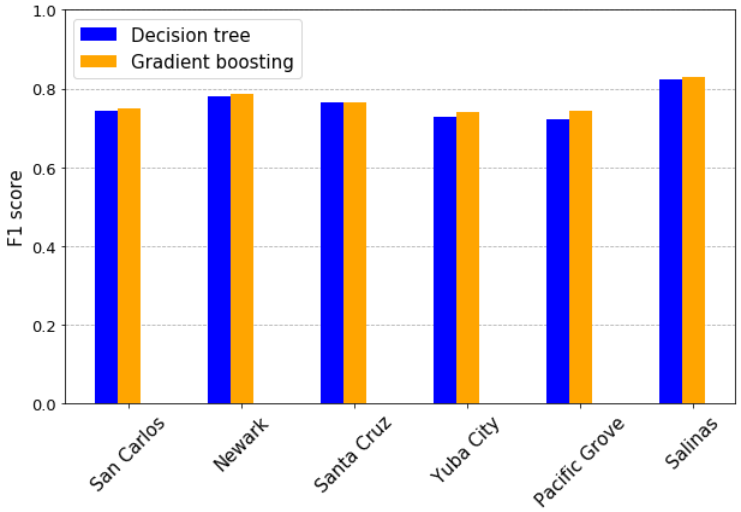

Figure 3.3: Comparison between two link prediction models: decision tree and gradient boosting. *The F1 score is the harmonic mean of precision and recall of link prediction (after matching equivalent segments) on the test areas in California.*

Test area	Precision	Recall	F1 score	# supplemented poles	Precision (after supplement)	Recall (after supplement)	F1 score (after supplement)
a. Test areas in California							
San Carlos, CA, U.S.A. (development set)	0.836	0.800	0.818	146	0.895	0.807	0.849
Newark, CA, U.S.A.	0.845	0.811	0.828	95	0.916	0.820	0.865
Santa Cruz, CA, U.S.A.	0.850	0.818	0.834	47	0.890	0.824	0.856
Yuba City, CA, U.S.A.	0.880	0.776	0.825	9	0.887	0.778	0.829
Pacific Grove, CA, U.S.A.	0.773	0.832	0.801	123	0.848	0.839	0.844
Salinas, CA, U.S.A.	0.816	0.943	0.875	73	0.888	0.939	0.913
Average (except San Carlos)	**0.833**	**0.836**	**0.833**	**69.4**	**0.886**	**0.840**	**0.861**
b. Test areas in Sub-Saharan Africa							
Ntinda, Kampala, Uganda	0.799	0.707	0.750	-	-	-	-
Kololo, Kampala, Uganda	0.853	0.717	0.779	-	-	-	-
Highridge, Nairobi, Kenya	0.895	0.647	0.751	-	-	-	-
Ngara, Nairobi, Kenya	0.887	0.584	0.704	-	-	-	-
Ikeja, Lagos, Nigeria	0.953	0.636	0.763	-	-	-	-
Average	**0.877**	**0.658**	**0.749**	-	-	-	-

Table T3.1: Pole localization performance on the test areas in California and Sub-Saharan Africa (SSA)*, with distance threshold* $D_{matching}$ *= 25m. F1 score is the harmonic mean of precision and recall.*

Test area	Precision	Recall	F1 score	# supplemented edges	Precision (after supplement)	Recall (after supplement)	F1 score (after supplement)
a. Test areas in California							
San Carlos, CA, U.S.A. (development set)	0.787	0.713	0.748	161	0.853	0.726	0.784
Newark, CA, U.S.A.	0.816	0.760	0.787	113	0.899	0.773	0.832
Santa Cruz, CA, U.S.A.	0.809	0.723	0.763	48	0.851	0.730	0.786
Yuba City, CA, U.S.A.	0.828	0.671	0.741	9	0.836	0.673	0.746
Pacific Grove, CA, U.S.A.	0.709	0.783	0.744	132	0.780	0.793	0.786
Salinas, CA, U.S.A.	0.774	0.891	0.828	85	0.856	0.889	0.872
Average (except San Carlos)	**0.790**	**0.730**	**0.757**	**92.6**	**0.844**	**0.739**	**0.787**
b. Test areas in Sub-Saharan Africa							
Ntinda, Kampala, Uganda	0.801	0.664	0.726	-	-	-	-
Kololo, Kampala, Uganda	0.826	0.639	0.721	-	-	-	-
Highridge, Nairobi, Kenya	0.879	0.701	0.780	-	-	-	-
Ngara, Nairobi, Kenya	0.793	0.510	0.621	-	-	-	-
Ikeja, Lagos, Nigeria	0.908	0.637	0.749	-	-	-	-
Average	**0.841**	**0.630**	**0.719**	-	-	-	-

Table T3.2: Link prediction performance (gradient boosting model) on the test areas in California and SSA. *F1 score is the harmonic mean of precision and recall (after matching equivalent segments).*

7.3.3. Subtask 3.3: Preliminary benchmark the performance of distribution grid GIS mapping algorithm

The goal of this subtask is to evaluate the overall performance of distribution grid GIS mapping model we developed by benchmarking its prediction against the ground truth distribution grid map (PG&E ICA map). It includes the evaluation for both overhead distribution lines and underground distribution lines. In this subsection, we first introduce the new method we developed for mapping underground lines, followed by the model evaluation metrics and results. Finally, we summarize the accomplishments for Milestone 3.3.1 and 3.3.2.

Undergrounding line mapping method: Street view images are only able to capture the information of overhead distribution grids. To estimate the grid map for areas where power lines are underground or street view images are not available, we develop a heuristic approach that integrates the information of the estimated overhead grid map, the road network, and the map of buildings for inferring underground grid map. A premise for this approach is that all buildings should be connected to grids, which

means that all buildings that are not connected by overhead grids should be connected by underground grids. Therefore, this approach is only applicable to regions with nearly 100% electrification rate. Under this premise, we estimate the underground grid map by first identifying buildings which have not been covered by the estimated overhead grid, and then running a modified Dijkstra's algorithm to generate paths to greedily connect all of them. The paths that are generated in this algorithm are used as the estimates of the underground grid.

To pick out unconnected buildings, we dilate the line connections of the estimated overhead grid with a radius R_{dilate} and overlay it with the building map. Buildings that are not covered by the dilated paths are treated as unconnected buildings. In the modified Dijkstra's algorithm, unconnected buildings are the targets that should be connected, and new paths are greedily generated on top of the estimated overhead grid until all targets are connected. The algorithm is run on the meshed spatial map where the overhead grid, roads, and buildings are discretized. Paths can be generated from one cell to any of its 8 neighbor cells (including diagonal neighbors). Weights of road cells are set to be lower than that of other cells. In this way, as the objective of the algorithm is to find the paths with minimum weights, connections following roads are more preferable. Such weight assignments are based on the grid construction practice that underground power lines are usually buried along roads to facilitate maintenance. The final output of underground grid inference is a 2D mask with binary values indicating whether each cell belongs to the underground grid or not.

Evaluation metrics: We evaluate the overall grid mapping performance on the meshed spatial map since the underground part of the grid cannot be explicitly represented as nodes and edges. To this end, both the ground truth map and the entire predicted grid map — including overhead and underground portions — are meshed into 2D binary arrays with the cell size 2m x 2m, denoted as G and H, respectively. Cells that belong to grids have value 1 and otherwise 0. To estimate the correct rate of the estimated grid map ("precision"), we dilate 1-value cells in G with a radius R_{eval} to generate G_{dilate}, then overlay G_{dilate} with H, and finally calculate the ratio of 1-value cells in H that can be covered by G_{dilate}. Similarly, to estimate the ratio of the ground truth grid map that can be detected within a distance ("recall"), we dilate 1-value cells in H with the same radius R_{eval} to generate H_{dilate}, then overlay H_{dilate} with G, and calculate the ratio of 1-value cells in G that can be covered by H_{dilate}. Hence the precision and recall for grid mapping are defined as:

$$precision\ of\ grid\ mapping = \frac{\left|G_{dilate} \cap H\right|}{|H|}$$

$$recall\ of\ grid\ mapping = \frac{\left|G \cap H_{dilate}\right|}{|G|}$$

Here $\cap$ means the intersection between two 2D binary masks, and | | means the number of 1-value cells in a binary mask.

Test area	Precision	Recall	F1 score
a. Test areas in California			
San Carlos, CA, U.S.A. (development set)	0.857	0.797	0.826
Newark, CA, U.S.A.	0.850	0.805	0.827
Santa Cruz, CA, U.S.A.	0.751	0.766	0.758
Yuba City, CA, U.S.A.	0.875	0.762	0.815
Pacific Grove, CA, U.S.A.	0.840	0.871	0.856
Salinas, CA, U.S.A.	0.924	0.926	0.925
Average (except San Carlos)	**0.848**	**0.826**	**0.836**
b. Test areas in Sub-Saharan Africa			
Ntinda, Kampala, Uganda	0.920	0.782	0.846
Kololo, Kampala, Uganda	0.962	0.782	0.863
Highridge, Nairobi, Kenya	0.971	0.802	0.878
Ngara, Nairobi, Kenya	0.982	0.655	0.786
Ikeja, Lagos, Nigeria	0.988	0.756	0.857
Average	**0.965**	**0.755**	**0.846**

Table T3.3: Overall grid mapping performance (using gradient boosting model for link prediction) on the test areas in California and Sub-Saharan Africa (SSA). *The performance is evaluated on the meshed spatial map using path dilation with dilation radius R_{eval} = 30m. Supplemented poles and line connections are considered.*

Milestone 3.3.1: Achieve both precision and recall for the similarity comparison between actual PG&E distribution grid map (above + underground) > 80%

We have achieved this milestone in Q3 of BP1. The evaluation metrics are detailed in subsection 7.3.3. We use R_{eval} = 30m as specified in the SOPO. The overall grid mapping performances including both overhead and underground grids are evaluated on the meshed spatial map of PG&E grid and the results are shown in Table T3.3. The gradient boosting is used as the link prediction model. As is shown, for the 5 test areas in Northern California, **83% - 97%** of the actual distribution grid can be detected within 30m ("recall"). For **89% - 98%** of the estimated distribution grid, actual distribution grids can be found within 30m. They are all higher than the target value **80%**.

Milestone 3.3.2: Demonstrate that the grid mapping method can identify and correct missing lines in the utility-owned grid maps

We have achieved this milestone in Q3 of BP1. Specifically, our framework can localize the poles that are not recorded in the PG&E ICA maps, and these newly detected poles can serve as supplements for the utility-owned data. We validate the presence of these supplemented poles by manually checking the Google street view images at their geolocations. The number of supplemented poles in 5 test areas ranges from **9 to 123** (Table T3.1, column 3). After considering the supplemented poles, the average recall of pole localization over 5 test areas is 0.886 (Table T3.1, column 4) and the average precision is 0.840 (Table T3.1, column 5). By identifying the missing poles, our model

can identify line connections that are not recorded in the ground truth dataset as supplements. The number of unrecorded connections in 5 test areas ranges from **9 to 132** (Table T3.2, column 3). After considering these unrecorded connections, the average recall of pole localization over 5 test areas is 0.844 (Table T3, column 4) and the average precision is 0.771 (Table T3.2, column 5). These results show that our proposed grid mapping method can identify **≥9 (target: >0)** poles/lines not documented in the utility-owned grid maps for each test area.

Region	Precision	Recall	F1 score
San Carlos	0.862	0.850	0.856
Newark	0.891	0.861	0.876
Santa Cruz	0.801	0.823	0.812
Yuba City	0.916	0.830	0.871
Pacific Grove	0.874	0.911	0.892
Salinas	0.940	0.947	0.944
Watsonville	0.850	0.856	0.853
Richmond	0.876	0.914	0.895
Livermore	0.851	0.888	0.869
Eureka	0.872	0.855	0.863

Table T3.4: Overall grid mapping performance benchmarked against PG&E data. *The performance is evaluated on a raster map using path dilation with a dilation radius R_{eval} = 30m. Here we define precision as the fraction of predicted distribution grid located within a distance Reval of ground truth grid, and define recall as the fraction of ground truth distribution grid that can be detected within a distance R_{eval}.*

7.3.4. Subtask 3.4: Run the distribution grid GIS mapping algorithm on data from multiple regions to produce a database of predicted GIS maps

The goal of this subtask is to deploy the distribution gird GIS mapping tool we developed to different regions in California and evaluate its generalizability to other parts of the world, especially countries in Sub-Saharan Africa where the electricity access is limited and the information about the electricity infrastructure is scarce.

Milestone 3.4.1: Deploy the grid GIS mapping model to no less than 10 areas

We have achieved this milestone in Q2 of BP2. In addition to the 8 regions included in the test set, we have deployed the distribution GIS mapping model to another two cities in California: Livermore and Eureka. The total number of regions we have deployed our

model on is **10 (target number: 10)**. Their performance benchmarked against PG&E data is listed in Table T3.4 (the evaluation metrics are detailed in subsection 7.3.3).

Specifically, to evaluate the generalizability of the distribution grid mapping framework, especially in developing countries where the energy infrastructure data are scarce, we transfer the framework developed using the data in California to 5 manually-curated test areas in Sub-Saharan Africa (SSA) and evaluate the model performance with the same metrics as defined in subsection 7.3.2 and 7.3.3 (i.e., precision and recall for pole localization, link prediction, and overall grid mapping, respectively). The 5 test areas are from three cities in SSA, including two areas in Kampala, Uganda, two areas in Nairobi, Kenya, and one area in Lagos, Nigeria. The World Bank maintains a geospatial dataset of transmission and distribution grids in Africa [27], but it only covers a few cities and most of the data in this dataset are for transmission lines. We correct errors in this dataset and identify additional overhead distribution lines by manually checking street view images and remote sensing images, and eventually construct the distribution grid maps for the 5 test areas in SSA that serve as the ground truth for model evaluation.

The line detector, the pole detector, and the link prediction model—detailed in subsection 7.3.1, 7.3.2 and 7.3.3—all remain the same without re-training or finetuning. All hyperparameters are also the same as those used in the California dataset except that the decision threshold to classify an image as positive is changed from 0.5 to 0.2 for the pole detector. Such a change is based on the observation that the utility poles in SSA are generally shorter than those in the U.S. which can make them more difficult to identify in upward street view images. Note that we do not predict the underground grid map for the SSA test areas since the assumption for underground grid mapping—all buildings are connected to the grid—does not necessarily hold in SSA, and the reference underground grid maps in SSA are not available for model evaluation.

Table T3.1 compares the pole localization performance between the California test areas and SSA test areas. While precisions of pole localization across the SSA test areas are generally higher than 0.8, the recall drops from an average of 0.84 in the California test areas to an average of 0.66 in SSA, which can be attributed to the difference in the appearance of utility poles between the US and SSA. Moreover, some utility poles in SSA are comparatively short, making them out of sight in upward street view images if they are not close enough to the locations where street view images were captured. Table T3.2 compares the link prediction performance between the California test areas (Table T3.2a) and SSA test areas (Table T3.2b). It shows that the link prediction recall drops from an average of 0.73 in the California test areas to an average of 0.63 in SSA (Table T3.2b). A potential mitigation approach is to augment the field of view (FoV) of upward images by leveraging the panoramic street views which are commonly captured in street view photography.

Table T3.3 compares the overall grid mapping performance between the California test areas (Table T3.3a) and SSA test areas (Table T3.3b). The framework achieves a precision from 0.92 to 0.99 and a recall from 0.66 to 0.80 in overall grid mapping (TTable T3.3b). The average precision increases from 0.85 on the California test areas

to 0.97 on the SSA test areas **(+12%)**. The average recall drops from 0.83 on the California test areas to 0.76 on the SSA test areas **(-7%)**. In Milestone 3.4.2, the target value of drop in precision and recall of grid mapping when transferring the model to regions outside the U.S. is set to be **<15%** (i.e., the target value is -15%). Therefore, we have achieved Milestone 3.4.2. This indicates that our framework, trained with the data in the U.S., can maintain a high correct rate and a reasonable detection rate of mapping when transferred to SSA even without re-training or fine-tuning.

7.3.5. Subtask 3.5: Extensive benchmark of the performance of distribution grid GIS mapping algorithm

The goal of this subtask is to evaluate the performance of our proposed grid mapping method on an alternative benchmark dataset to demonstrate the robustness of the model performance.

Milestone 3.5.1: Benchmark performance of distribution grid GIS mapping algorithm with detailed GIS maps provided by utility partners based on IAB advice

We have achieved this milestone in Q2 of the no-cost extension period. We have benchmarked the performance of the distribution grid mapping algorithm with PG&E's Electric Distribution GIS (EDGIS) dataset as suggested by PG&E and used the same overall grid mapping evaluation metrics as used in subtask 3.3 (detailed in subsection 7.3.3) to evaluate the model performance. The average recall and precision across the 5 test areas (the same as used in Subtask 3.2 and 3.3) are 85% and 90%, respectively. Both of them are at the same level as the ones in subtask 3.3 (detailed in subsection 7.3.3), indicating the robustness of the model performance.

7.4. Task 4: Visualization and applications development

In this task, the aim is to develop three applications around the Solar Energy Atlas database to demonstrate the value of the dataset produced by the project: (i) a browser-based tool that will allow visualizing the data produced and correlate it, at a minimum for the whole state of California, (ii) predictive and explanatory analysis on solar adoption that can potentially be used in planning and policy making, and (iii) a GIS application to correlate solar adoption with distribution grid characteristics, with a particular focus on wildfire resilience of distribution grids.

7.4.1. Subtask 4.1: Data browsing, correlation and visualization application

The goal of this subtask is to develop a browser-based platform to enable users to browse and visualize census-tract level aggregated data derived from the Solar Energy Atlas dataset.

Milestone 4.1: Integrate ≥ 5 data layers at aggregate level on the browser-based platform

We have achieved this milestone in Q3 of BP3. Specifically, we have developed the browser-based Energy Atlas platform at web.stanford.edu/group/energyatlas/home.html. The Energy Atlas platform provides an

"interactive map" module (web.stanford.edu/group/energyatlas/map.html) which multiple data layers:

1) Solar deployment rate (characterized by number of solar installations per 1000 households).
2) Solar radiation.
3) Demographic features, include:
    a. Average household income
    b. Average number of years of education
    c. Gini Index
    d. Population density
    e. Ratio of households that use coal/coke/wood as heating fuels
    f. Ratio of vacant housing units
    g. Ratio of owner-occupied housing units
    h. Ratio of family-occupied households
    i. Median housing unit value

at different aggregation levels (state, county, and census tract level). We also developed a "comparison" module (web.stanford.edu/group/energyatlas/dual-map.html) which enables users to correlate two different variables at the same level. The total number of layers that have been integrated is **11**, which is above the target number **5**.

7.4.2. Subtask 4.2: Identify the requirements for developing Energy Data Commons

The goal of this subtask is to identify the data types and classification for the Solar Energy Atlas and review the existing schema development process for Data Commons.

Milestone 4.2.1: Obtain a roadmap for the development of the schema for Energy Data Commons with a focus on Solar Energy Atlas

We have achieved this milestone in Q3 of BP3. Specifically, we have worked with the group leader (Dr. Ramanathan V. Guha, Google Fellow) and engineers in the Google Data Commons team to figure out the procedures for integrating our generated data into the Data Commons platform. The data import procedures follow the pipeline:

1. obtaining the source data
2. cleaning the data and representing it in the CSV format
3. Converting the data into one of Meta Content Framework (MCF), JSON-LD, or RDF format.

7.4.3. Subtask 4.3: Spatiotemporal pattern and underlying dynamics of solar adoption application

The goal of this subtask is to leverage Solar Energy Atlas to advance the understanding of the spatiotemporal pattern and underlying dynamics of solar adoption at a nationwide scale. This includes a correlational (explanatory) analysis to uncover and understand the socioeconomic factors that shape the spatiotemporal pattern of solar adoption; a causal analysis to identify the heterogenous effects of different types of solar energy

incentives; a predictive analysis to forecast solar installation growth at a spatially-resolved scale.

Correlational analysis: We conducted a correlation analysis of solar adoption by utilizing a technology diffusion model, called Bass model [28], to characterize the adoption trajectories from onset to saturation. Based on the Bass model, the solar adoption trajectory in each census block group can be segmented into four phases: pre-diffusion, ramp-up, ramp-down, and saturation (Figure 4.1A). Our results indicate that, by 2016, 55% block groups had not experienced any adoption at all while 15% had reached saturation (Figure 4.1B). The share of block groups that had already started adoption is consistently greater for higher income levels across time. For example, in 2016, 61% of high-income block groups had started adoption (Figure 4.1F) while only 30% of low-income block groups had (Figure 4.1C). However, among block groups that had started adoption already (Figure 4.2), 42% of the low-income block groups had entered the saturation phase in 2016, with a median saturation level of 2.8% (as a share of residential buildings). By contrast, only 30% of adopting high-income groups have saturated, at a median saturation level of 5.8%. This suggests that the PV adoption process in low-income communities is more likely to plateau yet at a lower adoption level given policy regime unchanged.

By correlating the Bass model parameters with socioeconomic variables (Figure 4.3), we find that block groups with higher income, higher education levels, higher PV benefit with rebate/grant, and lower percentage of renter-occupied housing units are more likely to have experienced an earlier onset of adoption. Compared to other communities, wealthier and more educated communities started adoption at lower levels of PV benefit with rebate/grant, implying that the PV benefit with rebate/grant is less relevant in high-income communities. However, we find that block groups with higher income levels actually have enjoyed higher PV benefit after the subsidization of rebate or grant (Figure 4.3C). This suggests that beyond low-income communities having comparatively lower adoption rates, for which there are many well-known contributing factors such as lower consumption capacity and higher fraction of renters, we find evidence that this lower adoption rate could be also related to the lower PV benefit with rebate/grant they experienced under a given incentive scheme. Given the inelasticity of high-income groups with respect to the PV benefit with rebate/grant we described earlier, our results suggest a potential for re-distribution of existing upfront subsidies to lower income communities to make PV adoption more equitable in its distribution.

We also find that the saturated adoption level is positively correlated with median household income, racial diversity, and average PV benefit with rebate/grant throughout 2006-2016 (Figure 4.3D), yet negatively correlated with the percentage of renter-occupied housing units. Interestingly, despite the correlation with earlier adoption onset, the level of education does not show a positive correlation (statistically significant) with saturated adoption level. This observation suggests a positive association of education levels in starting new adoption processes but not necessarily in increasing the eventually realized capacity.

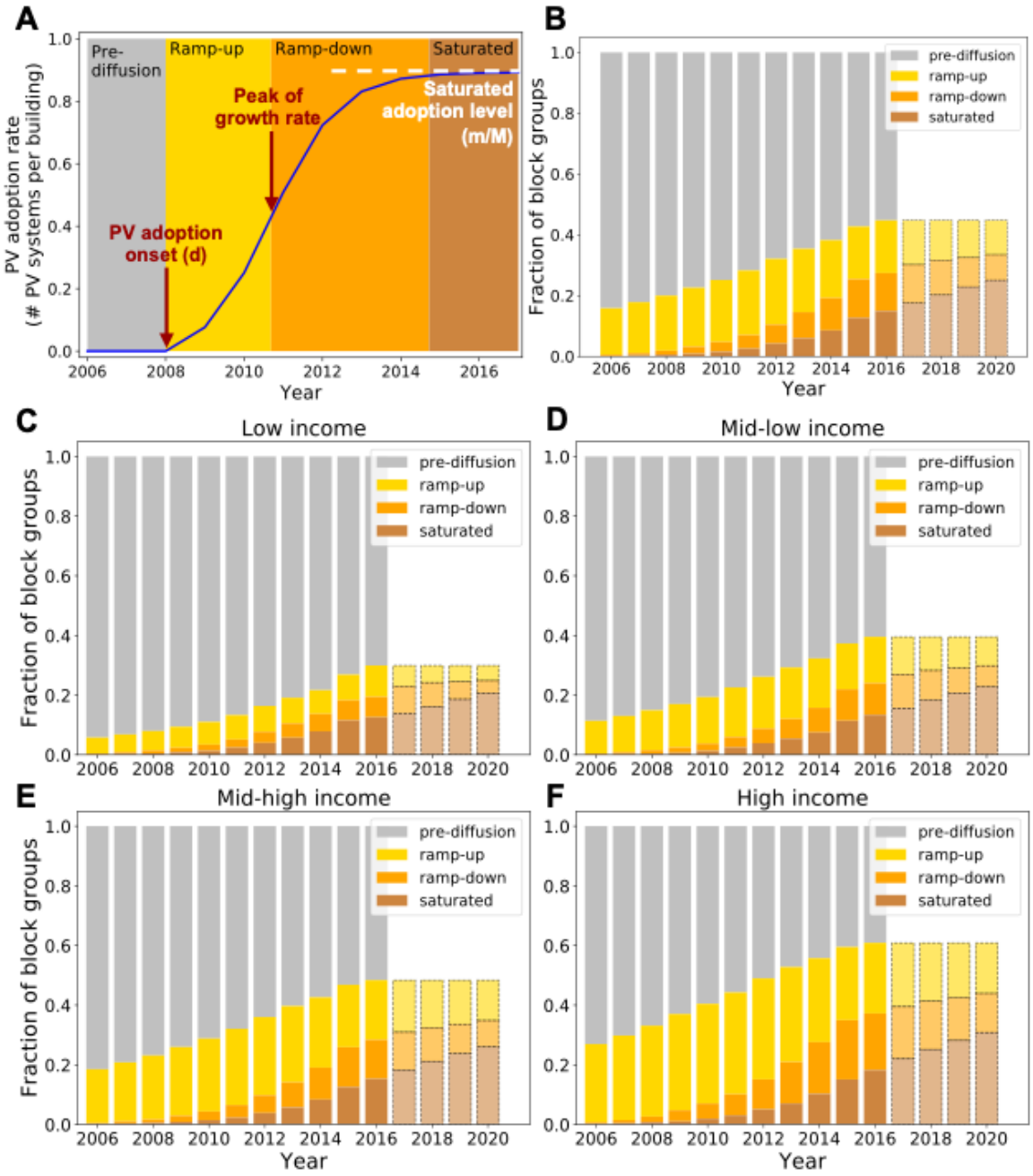

Figure 4.1: Four-phase segmentation of PV adoption trajectories. *(A) Illustration of the four phases of PV adoption according to the Bass model: pre-diffusion, ramp-up, ramp-down, and saturated. (B) Fractions of block groups in each of the four phases over time. Data from 2017 to 2020, marked with dashed edges, are projected by Bass models. No block groups are projected to exit the pre-diffusion phase and enter the ramp-up phase from 2017 onwards as we do not model the time when the adoption onset occurred. (C)-(F) Fractions of block groups in each of the four phases over time by income quartiles. Income quartiles are determined separately for each state.*

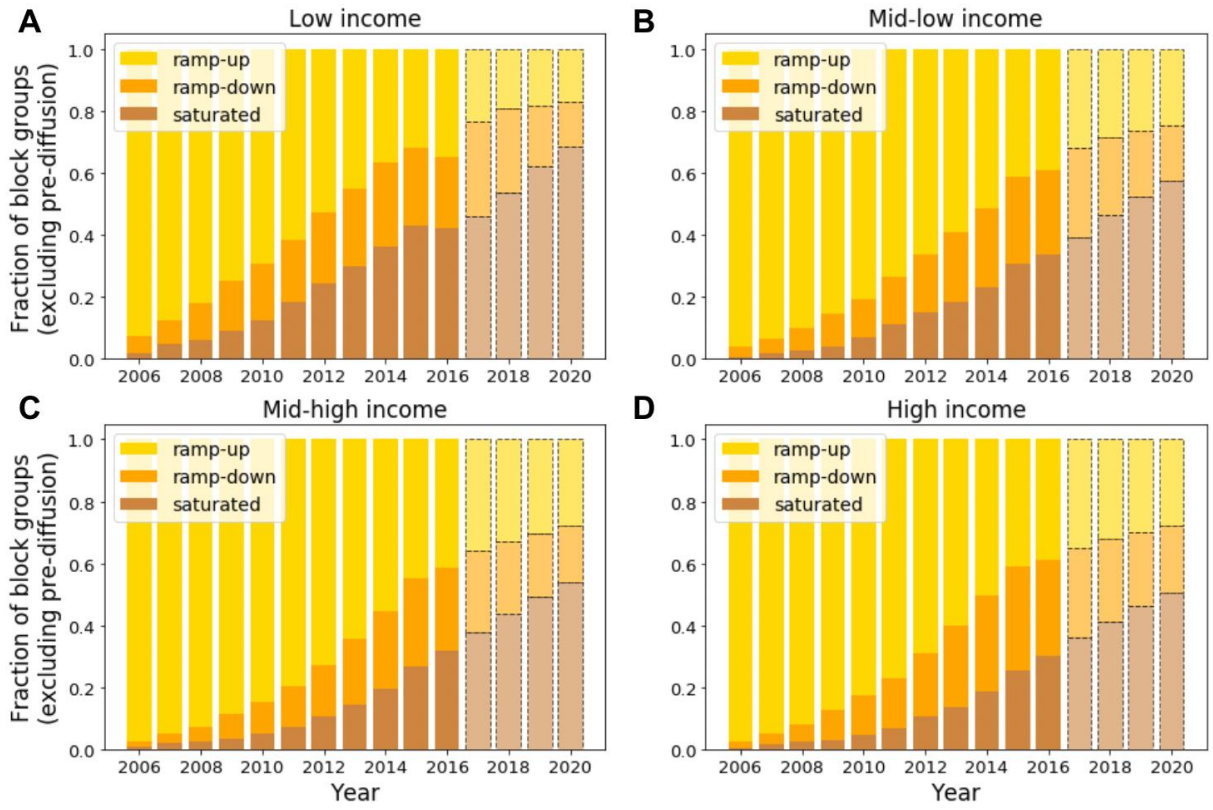

Figure 4.2: Fractions of block groups in "ramp-up", "ramp-down", and "saturated" phases ("pre-diffusion" phase is excluded from the base). *Values from 2017 to 2020, marked with dashed edges, are projected by Bass models. Among block groups that have already started the adoption process (i.e. not in "pre-diffusion" phase), lower-income block groups are more likely to be in "saturation".*

Causal analysis: we further conducted a causal analysis on the effects of solar incentives by identifying natural experiments of various incentives. Compared to randomized controlled trials (RCTs), a natural experiment study analyzes an event that is not under the control of researchers but naturally divides a population into exposed and unexposed groups to an intervention. Such a naturally occurring variation in exposure can be used to identify the effect of the intervention. To identify all natural experiments of incentive programs that were once present in the contiguous U.S., we build a spatiotemporal map of incentives based on the Database of State Incentives for Renewables & Efficiency (DSIRE). Each incentive program has a start date and potentially an end date. In the contiguous U.S., there are 994 incentive programs that are eligible for residential PVs. They can be divided into two major categories (financial incentives and regulatory policies) and further into 31 types (e.g., net metering, rebate, performance-based incentives). There are at most 31 incentive programs related to residential PV in a block group in a year.

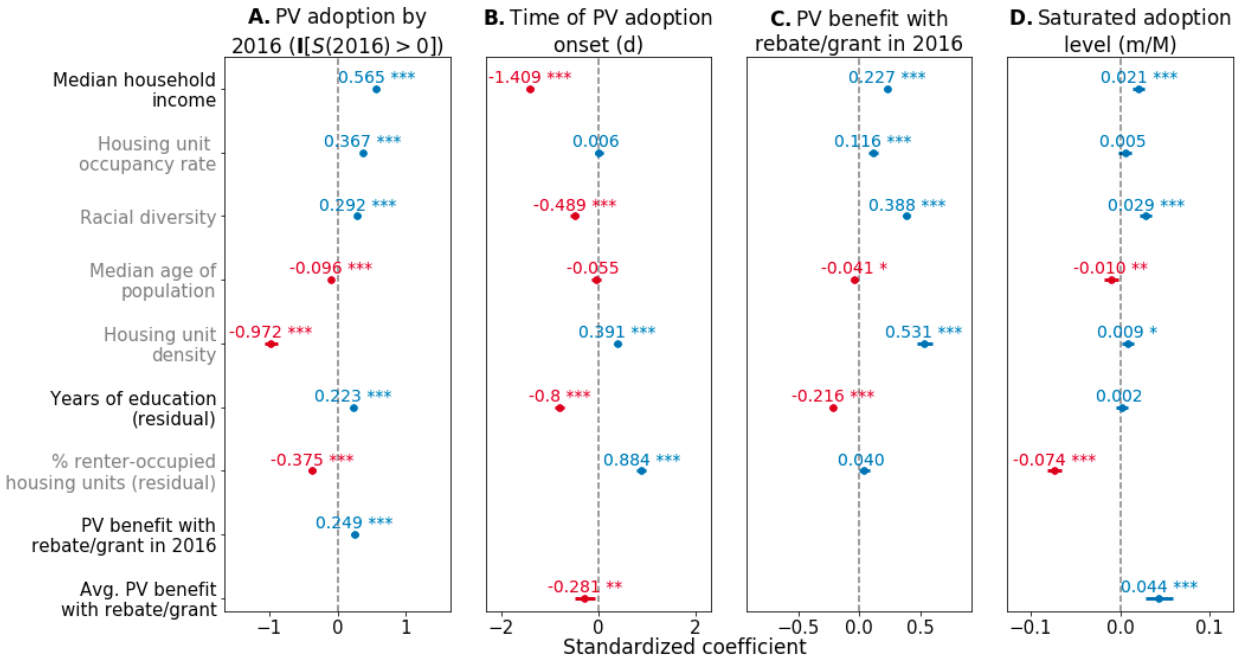

Figure 4.3: Standardized coefficients of regressions with demographic factors and PV benefit with rebate/grant. *Points represent coefficient estimates, bars 95% confidence intervals (CI), and statistical significance levels are denoted as *p<0.05, **p<0.01, and ***p<0.001. In each regression, state dummies are included to control for state-level variations but their coefficients are not shown here. Independent variables are normalized while dependent variables are not. PV benefit with rebate/grant, varying by location and time, is determined by residential electricity rate minus the Levelized Cost of Energy (LCOE) of residential PV after the subsidization of rebate or grant. We take the residual values of years of education and % renter-occupied housing units with respect to median household income to mitigate their mutual correlations. Racial diversity reflects the diversity of race and ethnicity in a block group. Other demographic variables, such as housing unit occupancy rate (percentage of housing units that are occupied), housing unit density (number of housing units per square mile), and percentage of renter-occupied housing units (percentage of housing units occupied by renters), are obtained from the American Community Survey (ACS) data.*
(A) Whether there has been PV adopted by 2016 vs. demographics and PV benefit with rebate/grant. Logit regression model is applied.
(B) The time of PV adoption onset d vs. demographics and PV benefit with rebate/grant.
(C) PV benefit with rebate/grant in 2016, characterized by USD cents/kWh, vs. demographics.
(D) Saturated PV adoption level vs. demographic characteristics and PV benefit with rebate/grant.

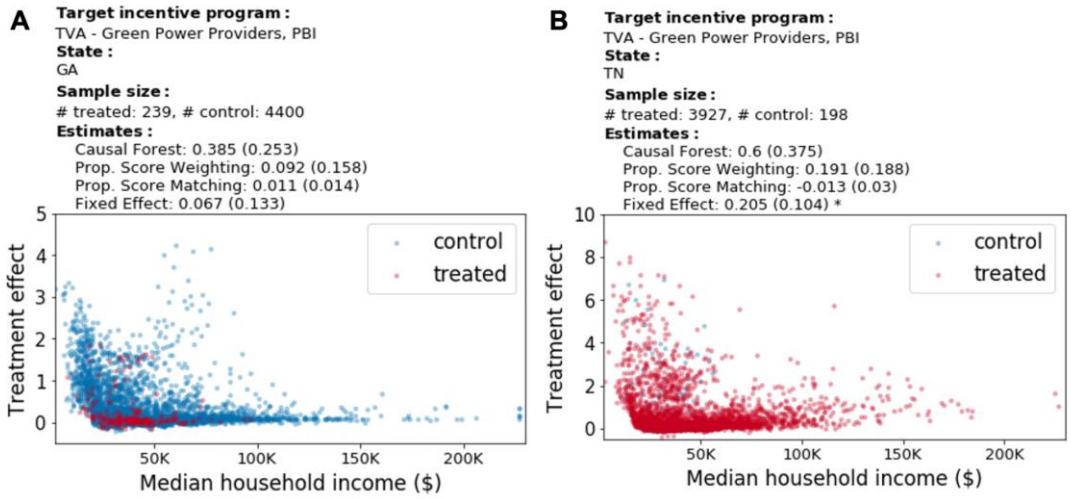

Figure 4.5: Treatment effect of performance-based incentive (PBI) programs vs. median household income, estimated by causal forest. *Each subfigure is corresponding to a single control-treatment group pair where the effect of the target incentive program is to be estimated. Each point is corresponding to a block group. "Estimates" show average treatment effects and their standard errors estimated by different models. Treatment effect is quantified in terms of number of installations per thousand households.*

Utilizing the spatiotemporal map of incentives, we can extract a "control-treatment group pair" for a specific incentive program, where both the control and treatment groups had the same set of incentives before year T except that a new incentive started in treatment group since year T while the incentives in control group still remained the same. This new incentive is called "target incentive" of which the effect can be estimated by existing causal inference models. We extract 170 control-treatment group pairs in total for residential PV incentives across the contiguous U.S.

For each pair, we apply a variety of causal inference models to identify the average treatment effect (ATE) of the target incentive on solar adoption rate, defined as number of solar installations per thousand households in a block group in a year. The models we apply include propensity score matching, propensity score weighting [29], and fixed effect regression. Moreover, causal forest [30] is a state-of-the-art causal inference model to estimate both ATE and heterogeneous treatment effect (HTE) by extending the random forest algorithm. We apply causal forest to each control- treatment group pair by using the difference of solar adoption rates between the year before the implementation of the target incentive and the first year after it. By using the fitted causal forest model to estimate the treatment effect for each block group, we can further obtain the correlation between the treatment effect and the median household income of each block group.

Figure 4.5 and Figure 4.6 show the results of control-treatment group pairs where we are able to observe the treatment effect at different income levels for performance-based incentives (PBI) and rebate programs, respectively. For each control-treatment group pair in these two figures, we show the ATE estimated by different models as well

as the correlation between treatment effects estimated by causal forest and median household income.

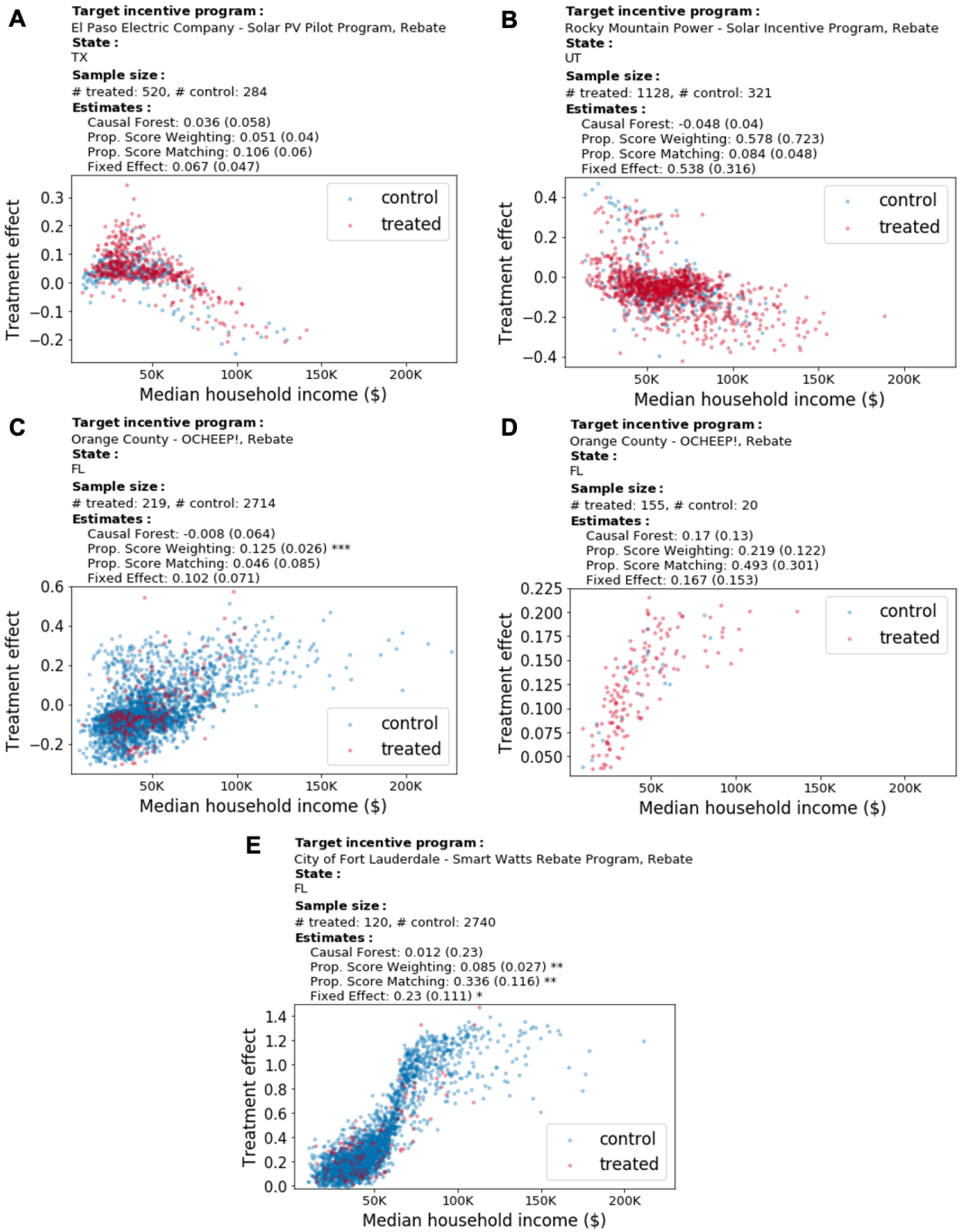

Figure 4.6: Treatment effect of rebate programs vs. median household income, estimated by causal forest. *Each subfigure is corresponding to a single control-treatment group pair where the effect of the target incentive program is to be estimated. Each point is corresponding to a block group. "Estimates" show average treatment effects (and their standard errors estimated by different models. Treatment effect is quantified in terms of number of installations per thousand households.*

For PBI, incentives are paid based on the actual energy production of the PV system. Typically, they are paid on the per kWh basis ($/kWh) over a period of time. Figure 4.5A and 4.5B show the effect of an PBI program, Green Power Providers offered by Tennessee Valley Authority (TVA), in two different states, Georgia (GA) and Tennessee (TN), respectively. We find that such a PBI incentive exhibits heterogeneous effects at different income levels. For block groups with median household income higher than $100K, the effect is close to 0 in Georgia and less than 2 installations per thousand households in Tennessee. By contrast, for block groups with income levels lower than $50K, a significant number of block groups experienced an effect of higher than 1 installation per thousand households in Georgia and 3 installations per thousand households in Tennessee. This suggests that this PBI program can activate solar adoption in low-income communities but is inactive in high-income ones.

For rebate programs, incentives are paid based on the installed capacity of the PV system to reduce its upfront cost. Typically, they are paid on the per kW basis ($/kW) (contrary to $/kWh in PBI). Figure 4.6 shows the effects of different rebate programs and their variations with income levels. We find that, for the two rebate programs displayed in Figures 4.6A and 4.6B (Solar PV pilot program offered by El Paso Electric Company, and Solar Incentive Program offered by Rocky Mountain Power), the effects are stronger in many low-income communities and show a slightly negative correlation with income levels. By contrast, for the two rebate programs displayed in Figures 4.6C to 4.6E (OCHEEP! offered by Orange County in Florida, and Smart Watts Rebate Program offered by the city of Fort Lauderdale), the effects show a positive correlation with income levels. We notice that one difference in incentive magnitude is that: for the former two programs, the maximum incentive amount, i.e., the upper bound of rebate that can be obtained by a customer, is $7,500 for the El Paso solar PV pilot program (Figure 4.6A) and $4,600 for the Rocky Mountain Power solar incentive program (4.6B). By contrast, the last two rebate programs both have a maximum incentive amount of only $1,000 (OCHEEP! in Figure 4.6C and 4.6D, and Smart Watts Rebate program in Figure 4.6E). Low maximum incentive amounts, therefore, may be a potential reason explaining why the effects of rebate programs in 4.6C to 4.6D are lower in lower-income communities, as a rebate program might need to have a high upper bound of cost deduction to be appealing to low-income communities.

To summarize, our results suggest that PBI can activate solar adoption in low-income communities while rebate programs have such effects only if the incentive amount is significant.

Milestone 4.3.1: Use regression model and solar installation data to predict solar adoption patterns considering spatial heterogeneity and achieved an out-of-sample regression $R^2 > 0.5$

We have achieved this milestone in Q3 of BP2. Specifically, we developed a predictive model based on random forest to forecast future solar installations in a local neighborhood given installations in previous years along with demographic features and incentives. The out-of-sample R^2 in predicting solar installations in the next year can

achieve **0.65** (target value: **0.5**). This model can be used for forecasting future solar installation growth at a granular geographic level to facilitate grid planning.

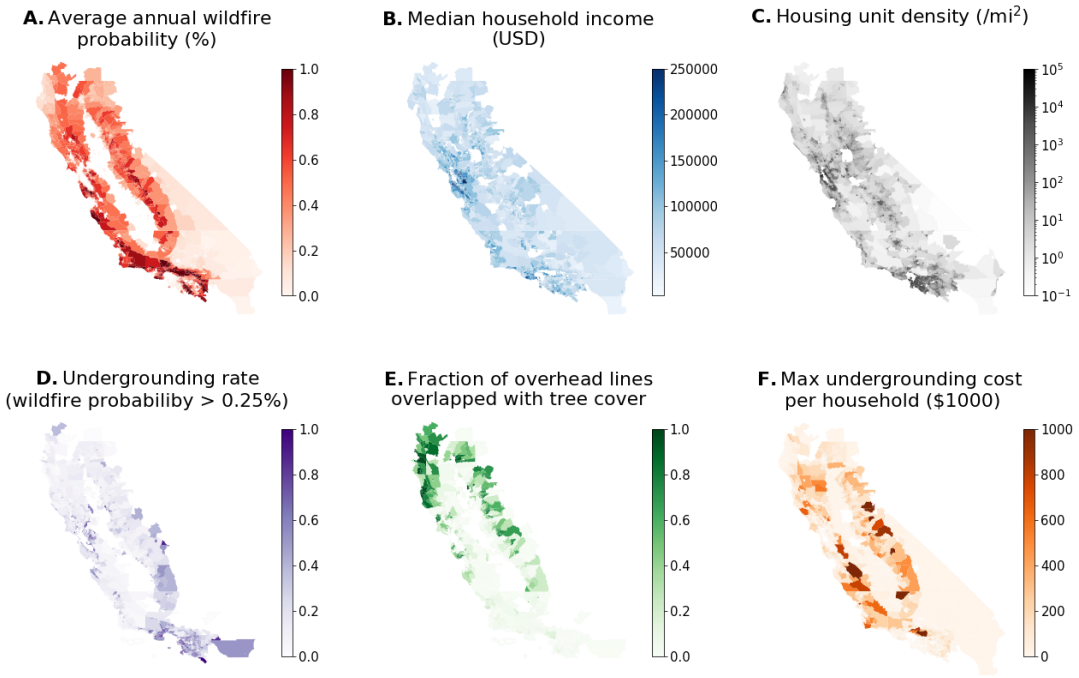

Figure 4.7: Geospatial distributions of wildfire probability, demographic characteristics, and distribution grid characteristics at the census block group level. *(A) Average annual wildfire probability over 2026-2050. (B) Median annual household income. (C) Housing unit density. (D) Undergrounding rate for distribution lines with annual wildfire probability > 0.25%. Undergrounding rate is defined as the fraction of distribution power lines buried underground in terms of length in a block group. (E) The fraction of overhead power lines that are overlapped with tree cover with tree height > 10m, which is used to characterize the proximity of power lines to trees. The tree cover map is at a spatial resolution of 10m. (F) Maximum undergrounding cost per household in each block group, under the scenario that overhead lines with wildfire probability > 0.25% are to be undergrounded and the cost is only shared locally within each block group.*

7.4.4. Subtask 4.4: Correlating solar power locations and overhead lines in distribution grids

The goal of this subtask is to combine the distribution grid GIS mapping with the produced solar installation data to uncover the distribution grid vulnerability to wildfires. This subtask can provide new insights into how grid adaptation approaches (e.g., undergrounding) and solar PV preparedness differ across different communities, and how to reduce the inequity in the wildfire resilience of distribution grids.

Milestone 4.4.1: Obtain overhead line ratio and solar PV capacity for 100% of very-high-fire-risk regions in PG&E territory

We have achieved this milestone in Q1 of BP2. We utilize the geospatial data of distribution grids of California's two major utilities—PG&E and Southern California Edison (SCE), and overlay them with the maps of tree cover [31] and predictive wildfire

probability over 2026-2050 [32, 33]. We further use the power line detection model developed in Task 3 (section 7.3) to estimate the fraction of distribution power lines in each block group that are buried underground in terms of line length, denoted as "undergrounding rate" (i.e., 1 – overhead line ratio). As a result, we have obtained the undergrounding rate and overhead line ratio for **100% very-high-fire-risk regions** (annual wildfire risk probability > 0.5%) and **100% high-fire-risk regions** (annual wildfire risk probability > 0.25%) for both PG&E and SCE territories. This has satisfied and gone beyond the **targeted coverage 100%** (for only very-high-fire-risk regions in PG&E territory)**.**

Utilizing such data, we further analyze the status quo of distribution grid vulnerability to wildfires in California at the census block group level—a highly granular geographic aggregation defined by the US Census Bureau. Figure 4.7 shows the geospatial distributions of wildfire probability, demographic characteristics, and various distribution grid characteristics in California.

We plotted the correlations between median annual household income and various distribution grid characteristics conditioning on wildfire threat—characterized by average annual wildfire probability over 2026-2050—for PG&E, SCE, and both territories (see Figure 4.8). We find that, conditioning on wildfire threat, undergrounding rates are positively correlated with median household income in both PG&E and SCE territories For high-fire-threat block groups (annual wildfire probability of distribution lines > 0.74%), the undergrounding rate is expected to be 65% at the income level of $200K but only 34% at $50K. Undergrounding rates in SCE territory are generally higher than PG&E territory, especially in high-fire-threat and low-income areas, which might result from the differences in undergrounding cost, ages of neighborhoods, constructability of underground lines, etc.

Apart from undergrounding rates, we further investigate the vulnerability of the overhead part of the grids across different communities. By overlaying the tree canopy map with that of distribution grids, for each block group we estimate the fraction of overhead power lines that are overlapped with tree canopy cover to represent the potential exposure of overhead lines to nearby vegetation (see its geospatial distribution in Figure 4.7E). Trees shorter than a threshold of 10m are filtered out as they are not likely to impact grids, according to the typical heights of poles and lines. We find that (Figure 4.8), in mid- and high-fire-threat areas of PG&E territory, lower-income block groups tend to have higher fractions of overhead lines overlapped with tree cover, indicating higher exposure of their grids to vegetation. However, such fractions are significantly lower in SCE territory. This may be explained by the sparser tree cover and lower tree heights in SCE territory (Southern California) than that PG&E territory (Northern California).

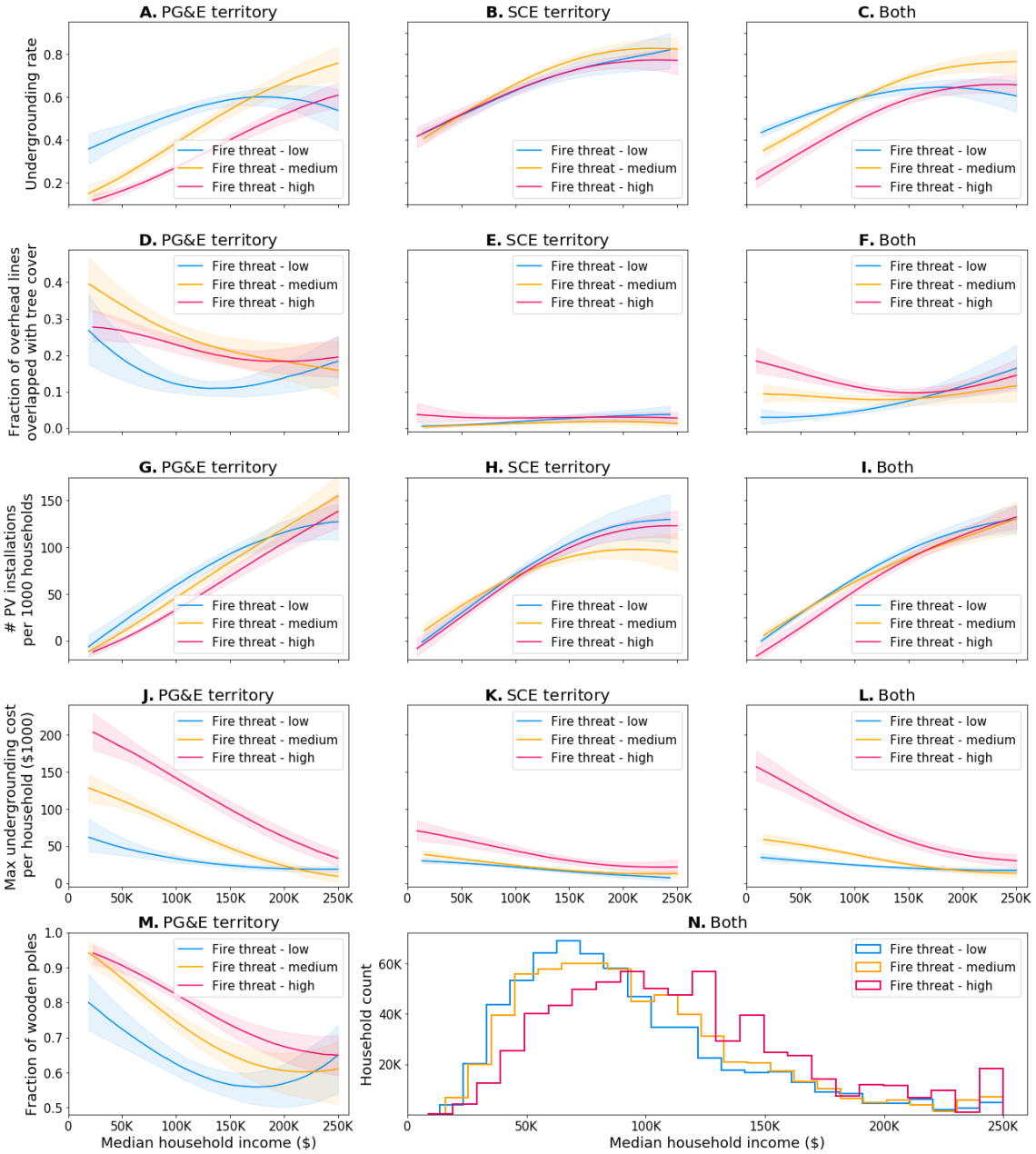

Figure 4.8. Correlations between distribution grid characteristics and median household income conditioning on wildfire threats in PG&E territory (left column), SCE territory (middle column), and both (right column). *Curves are fitted using locally weighted scatterplot smoothing (LOWESS). Lighter areas represent 90% confidence intervals (CIs) obtained by 1,000 bootstraps of curve fitting. Wildfire threat stratification is based on the tertiles of maximum wildfire probability of distribution lines in a block group. Dependent variables include: (A)-(C) undergrounding rate, (D)-(F) the fraction of overhead power lines that are overlapped with tree cover (tree height > 10m), (G)-(I) number of residential PV installations per 1,000 households, (J)-(L) maximum undergrounding cost per household under the scenario that overhead lines with wildfire probability > 0.25% are to be undergrounded and the cost is only shared locally within each block group, and (M) the fraction of wooden utility poles (data only available for PG&E). (N) Household count in each bin of median household income, conditioning on wildfire threats.*

Wooden poles are more vulnerable to fires and to vegetation strike than poles made of other materials such as concrete or steel. Utilizing PG&E's grid asset data, we find that lower-income block groups in PG&E territory tend to have higher fractions of wooden poles conditioning on wildfire threat (Figure 4.8). This suggests that besides lower undergrounding rates, low-income communities are also less likely to install more fire-resistant poles (e.g., steel, concrete) which may be attributed to their higher cost than wooden ones.

If the grid itself is vulnerable, preemptive de-energization is likely to be taken as the last resort option to prevent wildfires. Standalone power systems or grid sectionalization relying on locally-sited DERs, such as solar photovoltaics (PV) and batteries, are also considered as an alternative approach to provide electricity for high-fire-threat areas to prevent wildfires. We find that, however, solar PV adoption rates, characterized by the number of PV installations per thousand households, are lower in low-income communities at different wildfire threat levels (Figure 4.8). This suggests that, apart from less undergrounding protection and higher vulnerability of the overhead part of the grids, electricity, especially the one provided by renewable DERs, is also less accessible to low-income communities when they have to be disconnected from major grids as the last resort option facing wildfire threats. Note that grid-tied PVs need to be accompanied by batteries to provide power during outage, but battery adoption is not analyzed here.

Overall, low-income communities not only have less undergrounding protection of distribution lines, but also have higher vulnerability of the overhead part of grid infrastructure to wildfires and less DER preparedness for last-resort wildfire prevention approaches.

7.4.5. Subtask 4.5: Develop the Data Commons schema for Solar Energy Atlas

The goal of this subtask is to develop and implement schema.org schema for Solar Energy Atlas following the Data Commons development process and validate the schema with Data Commons.

Milestone 4.5.1: Validate deployment of Solar Energy Atlas on Data Commons to ensure 0 mismatch compared with results obtained from original offline data within numeric tolerance

We have achieved this milestone in Q3 of BP2. Specifically, we have implemented the schema for the census-block-group-level time-series solar installation data produced in Task 1. The Schema is implemented with the Meta Content Framework (MCF) format. Specifically, for the above solar installation data frame, each row contains the data from each census block group indexed by its FIPS code "blockgroup_FIPS". There is a column named "cumulative_num_of_residential_PVs_by_[X]" for each year X, where [X] is a placeholder for each year from 2005 to 2017. The MCF file looks like follows:

```
Node: E:data->E0
typeOf: dcs:StatVarObservation
```

```
observationAbout: C:data->blockgroup_FIPS
observationDate: 2005
variableMeasured: Number_of_residential_PVs
value: C:data->cumulative_num_of_residential_PVs_by_2005

Node: E:data->E1
typeOf: dcs:StatVarObservation
observationAbout: C:data->blockgroup_FIPS
observationDate: 2006
variableMeasured: Number_of_residential_PVs
value: C:data->cumulative_num_of_residential_PVs_by_2006

Node: E:data->E2
typeOf: dcs:StatVarObservation
observationAbout: C:data->blockgroup_FIPS
observationDate: 2007
variableMeasured: Number_of_residential_PVs
value: C:data->cumulative_num_of_residential_PVs_by_2007
```

Here we only show the schema for year 2005-2007, while the remaining years (2007-2017) follow the same way. Using the same approach, we have also implemented the schema for census-block-group data on California's distribution grid undergrounding status (characterized by the fraction of power lines buried underground) produced by the grid mapping framework. Comparing the correlational analysis results (see details in subsection 7.4.3 and 7.4.4) using online data vs. using offline data, we verify that they have **zero mismatch (target: zero mismatch)**.

7.4.6. Subtask 4.6: Upload the Solar Atlas data to Data Commons and test integration

Milestone 4.6.1: Upload Solar Energy Atlas data to Data Commons

We have achieved this milestone in Q3 of BP2. Specifically, we have uploaded the census-block-group-level time-series solar installation data as well as the distribution grid data introduced above together with their Schema (.tmcf file) to the Data Commons platform to make them publicly available. Our tests show that there is no mismatch between the correlations derived with online data and the ones obtained from original offline data (see Milestone 4.5.1). Figure 4.9 shows an example of the residential solar PV installation curve for a block group in Miami-Dade County, Florida.

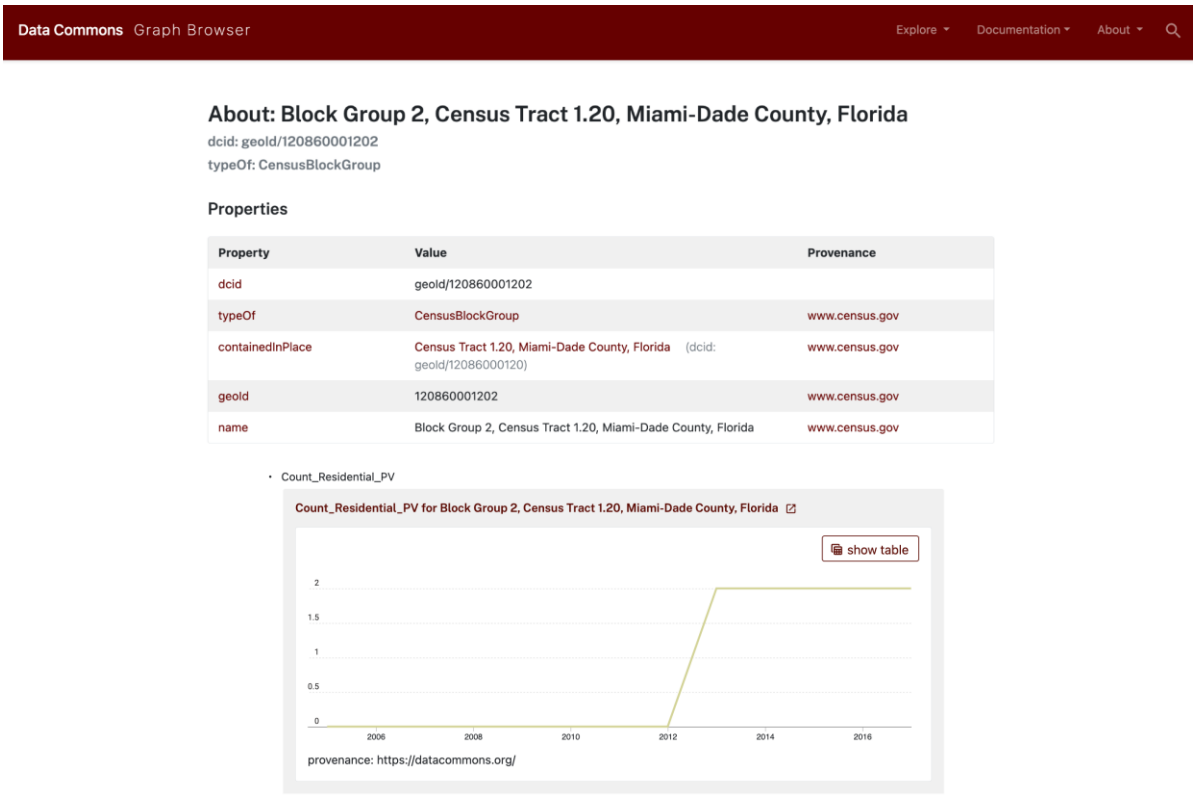

Figure 4.9: A visualization of the residential solar PV installation curve of block group #120860001202 in Miami-Dade County, Florida on the Data Commons platform. *This data is imported from the DeepSolar++ dataset produced in Task 1.*

7.5. Task 5: Data Management and technical advisory board

In this task, we aim to convene meetings with industry advisory board to collect their interested use cases and the data that can be shared in each budget period. We also aim convene a webinar in each budget period with experts from industry and academia to discuss applications and technical challenges in applying machine learning for renewable energy and grid research.

7.5.1. Subtask 5.1: Convene industry advisory board meeting for Year 1

Milestone 5.1.1: Convene TAC meeting in Year 1 to identify potential use cases and seek high-quality non-public input data

We have achieved this milestone in Q3 of BP1. We have formed a Technical Advisory Committee (TAC) (formerly called Industry Advisory Board). There are 7 industry members and most of them have current or prior experience working in a utility company. A TAC meeting was held (via zoom video conference) on 2022/04/22. The names of the people attending this meeting are listed below. All TAC members attended except one, who is very interested in the project and data and will attend future meetings. Below we summarize some of the key discussion.

TAC members: (1) Jonathan Bradshaw, Pacific Gas & Electric, (2) Andy Eiden, Portland General Electric, (3) Luke Forster, NYSERDA, (4) Amir Kavousian, Altitude Networks, (5) Lena Perkins, City of Palo Alto Utilities, (6) Diego Ponce, East Bay Community Energy, (7) Liuxi (Calvin) Zhang, Eaton

Stanford DeepSolar Team members: Ram Rajagopal, June Flora, Tiffany Branum, Chin-Woo Tan, Chad Zanocco, Zhecheng Wang, Rajanie Prabha

Welcome & Overview	Ram Rajagopal
Introduction of Members	Chin-Woo Tan
Previous DeepSolar Research Current Research Plans & Work to Date	Zhecheng Wang
Questions & Discussion	ALL
Updating the 2017 DeepSolar Database	Rajanie Prabha
Projected applications of the DeepSolar Database	Chad Zanocco
Questions & Discussion	ALL

Table 5.1: DeepSolar TAC meeting agenda, 2022-04-22

Overview of TAC meeting: Members were very engaged in the presentation and Q&A with discussion taking up the remaining time after the presentation by PhD student Z. Wang. In order to give members a complete understanding of the DeepSolar project, we video recorded the last two presentations, and placed videos and accompanying slides in a shared drive and sent those materials all members.

Topics of Discussion: An initial categorization of the topics of discussion are (1) grid resiliency, (2) meeting the electricity service needs of underserved populations, (3) discovering the potential for behind the meter (BTM) resources - future solar, battery installation, EV purchase and identifying micro-grid, commercial, and community solar sites. Finally, there was a wide-ranging discussion regarding the use of Diffusion of Innovation Theory and its application to predicting current growth curves of solar and perhaps other BTM resources.

7.5.2. Subtask 5.2: Convene a webinar on computer vision applications for grid

Milestone 5.2.1: Convene a webinar in Year 1 on computer vision applications for grid

We have achieved this milestone in Q4 of BP1. The webinar on computer vision applications in power grid management was held on August 15, 2022. We have invited the founder and CTO of Buzz Solutions, Vikhyat Chaudhry, as a guest speaker, and grid experts from different companies as an advisory committee. The webinar started

with three presentations followed by an open discussion session on the potential use cases of computer vision in a variety of power system applications, from resource mapping to power system fault detection.

7.5.3. Subtask 5.3: Convene industry advisory board meeting for Year 2

Milestone 5.3.1: Convene IAB meeting in Year 2 to demonstrate initial datasets for Solar Energy Atlas

We have achieved this milestone in Q3 of BP2. We invited members of our Technical Advisory Committee (TAC) (formerly called Industry Advisory Board) to this meeting on 2023/04/06. There are 7 industry members and most of them have current or prior experience working in a utility company. The names of the people invited to this meeting are listed below. All TAC members attended except one, who is very interested in the project and data and will attend future meetings. Below we summarize some of the key discussion.

TAC members: (1) Jonathan Bradshaw, Pacific Gas & Electric, (2) Andy Eiden, Portland General Electric, (3) Luke Forster, NYSERDA, (4) Amir Kavousian, Altitude Networks, (5) Lena Perkins, City of Palo Alto Utilities, (6) Diego Ponce, East Bay Community Energy, (7) Liuxi (Calvin) Zhang, Eaton

Stanford DeepSolar Team members: Ram Rajagopal, June Flora, Chad Zanocco, Zhecheng Wang, Rajanie Prabha, Chin-Woo Tan, Tiffany Branum.

Discussion: Stanford post-doc Chad Zanocco started with a quick update and overview of the DeepSolar Energy Atlas project and their applications. The presentation was followed by discussion on various related topics, including incentives to install solar especially for disadvantaged communities and low-income households, fire code for solar installation, grid resiliency to wildfires, and the value of crowdsourced power inspection.

7.5.4. Subtask 5.4*:* Convene webinar in Year 2 to discuss application using Solar Energy Atlas and data sharing

Milestone 5.4.1: Convene IAB meeting in Year 2 to demonstrate initial datasets for Solar Energy Atlas

We have achieved this milestone in Q2 of the no-cost extension period. In this subtask, the goal is to convene a webinar on the potential use cases of Solar Energy Atlas. We invited four members from our Technical Advisory Committee (TAC) to this webinar on 3/13/2024. They have current or prior experience working in a utility company. The names of the people invited to this meeting are listed below. Below we summarize some of the key discussion.

- **TAC members**: (1) Jonathan Bradshaw, Pacific Gas & Electric, (2) Andy Eiden, Portland General Electric, (3) Diego Ponce, East Bay Community Energy, (4) Jorge Meraz, Pacific Gas & Electric

- **Stanford DeepSolar Team members**: June Flora, Chad Zanocco, Zhecheng Wang, Rajanie Prabha, Chin-Woo Tan

Discussion: Stanford post-doctoral researcher Zhecheng Wang started the presentation with an overview of the DeepSolar Energy Atlas project and its practical applications. Another segment of the webinar, covered by PhD student Rajanie Prabha, featured updates regarding the new DeepSolar database. Notably, the database now encompasses all solar installations up to the year 2022. The pipeline is updated with more robust vision transformer models to continually update data in the subsequent years. The final part of the presentation, covered by post-doctoral researcher Chad Zanocco, talked about the application of the DeepSolar database in identifying the non-residential equity gap, soon to appear in Nature Energy.

Jorge, from PG&E, raised a discussion around understanding this even distribution of PV installations across block groups at different geographical levels such as state, county, and city. Jon, another PG&E attendee, expressed interest in obtaining more details about obtaining specific values for the Bass model parameters, used by the DeepSolar timelapse to characterize the adoption trajectories from onset to saturation. Various attendees suggested the need for an API that can be queried for specific information, instead of downloading the full dataset. Furthermore, there is an added interest for solar adoption forecasting models predicting, where spatially, new PV installations might appear in the future. The webinar ended with the emphasis on the potential policy implications of utilizing DeepSolar data for targeted "repurposed energy" deployments, such as community-scale solar projects on underused industrial lands, alongside the importance of addressing knowledge gaps in non-residential solar modeling.

7.6. Additional task: Updating the DeepSolar database

The original DeepSolar database only documented solar installations as of 2017. Thus, it cannot provide updated information about solar adoption in recent years. In this additional task, we aim to fill this gap by developing more cost-effective and efficient data collection and machine learning methods to update the DeepSolar database using satellite images in recent years.

As a result of this subtask, we have generated the latest DeepSolar database as of 2023 with 2.95 million solar systems all across the country. To do so, we revamped the data acquisition pipeline for the DeepSolar dataset in three phases summarized as follows (Phase I, Phase II and Phase III).

Phase I: We used Microsoft Maps dataset to get country-wide open building footprints of the United States. Using these coordinates, we downloaded around 230 million image tiles across the US via Google Maps khms API. The downloading process took around six months. A few samples of the data are shown in Figure 6.1.

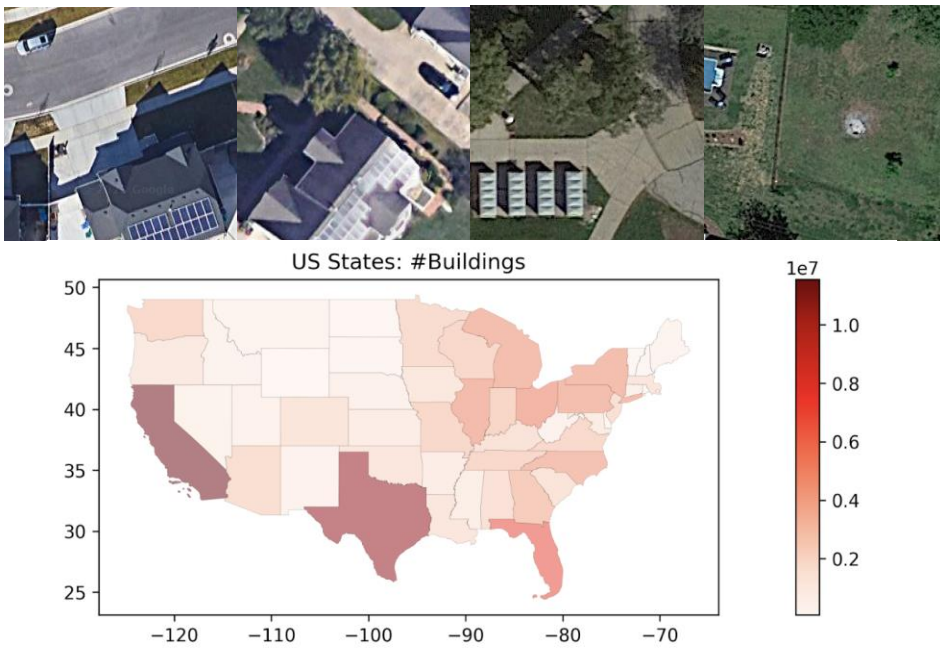

Figure 6.1: A few samples of image tiles collected (above). Data collected distribution across the US (below).

Phase II: In this phase, we identified positive image tiles within the acquired dataset. To achieve this, a variant of the Vision Transformer, ViTMAE (Masked Autoencoders), model was employed, which was fine-tuned on a limited dataset consisting of only 45,000 training and 5,000 validation labels. Various fine-tuning methods were explored with LoRA (Low-Rank Adaptation) outperforming all other strategies (Figure 6.2). It is noteworthy that this dataset constitutes less than 15% of the data labels used for training the Inception net model in the context of the 2017 DeepSolar project. The training is conducted within the framework of a supervised learning binary classification task, where the objective is to discriminate between the presence and absence of photovoltaic (PV) installations in the image tile. Subsequently, the model's performance is subsequently assessed using the gold standard evaluation dataset, which encompasses 92,000 image labels spread across the US. The model's test set Precision is 0.94 and recall is 0.91. The model is deployed on all the image tiles collected to get positive detections all across the US. It is important to note that one PV installation can be spread across many tiles, so the number of image tiles is not equal to the number of PV installations.

Once we have all the positive samples (PV-detected tiles), we use another model, Segformer, to get the solar PV segmentation boundary from the image tile. The Segformer model is trained with 5,607 supervised training labels, validated with 300 labels, and tested on 600 labels with a mean IOU of 0.92 and PV segmentation class IOU of 0.86 on the test set. After deploying this model on all the positively detected images, we get the segmentation mask for each image tile. Some of the prediction results are shown in Figure 6.3.

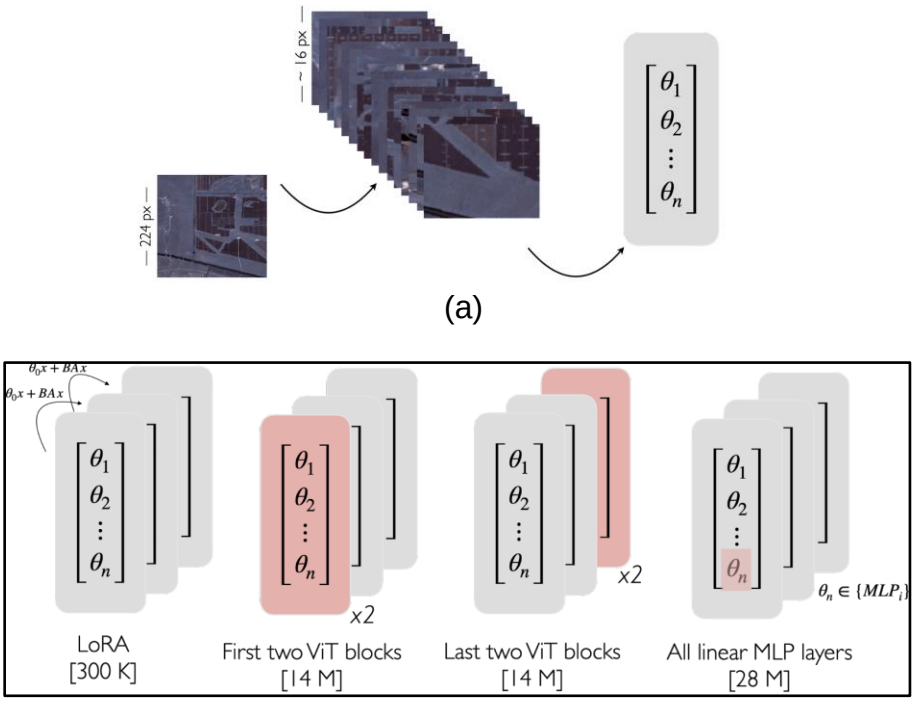

Figure 6.2: VitMAE model fine-tuning for binary classification task. *(a) VitMAE fine tuning. The image tile is presented to the model as a sequence of fixed-size patches and the parameters θ are fine tuned. (b) Fine-tuning strategies: LoRA (Best), Fine Tuning only the first two transformer blocks, finetuning only the last two transformer blocks, fine tuning all linear layers (left to right).*

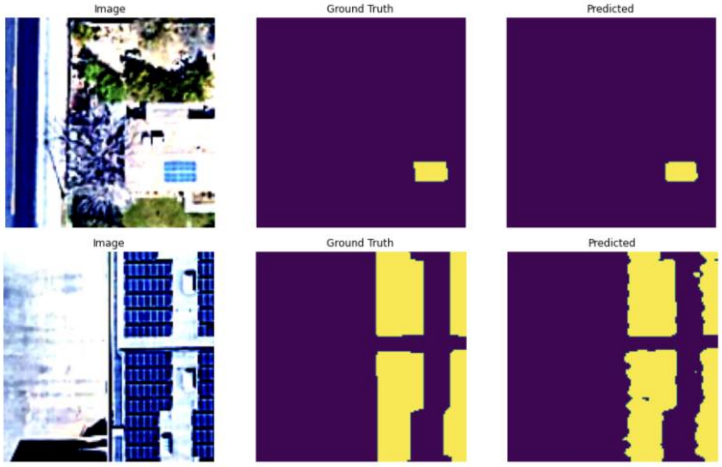

Figure 6.3: Segformer prediction results. *Columns show original images, ground truth annotation, and predicted mask of solar panel.*

Phase III: The final phase involves merging the image tiles so that a Solar PV split across multiple image tiles can be combined as one solar system. We identified around 2.95 million PV installations across the US as of 2023. Simultaneously, we also classify the PV systems into various categories: residential, commercial, utility, and solar heat for better insights into the deployment of residential and non-residential PVs. We use the ResNet 50 model, also trained in a supervised fashion to predict the category. Figure 6.4 shows the additional PV systems adopted across the US.

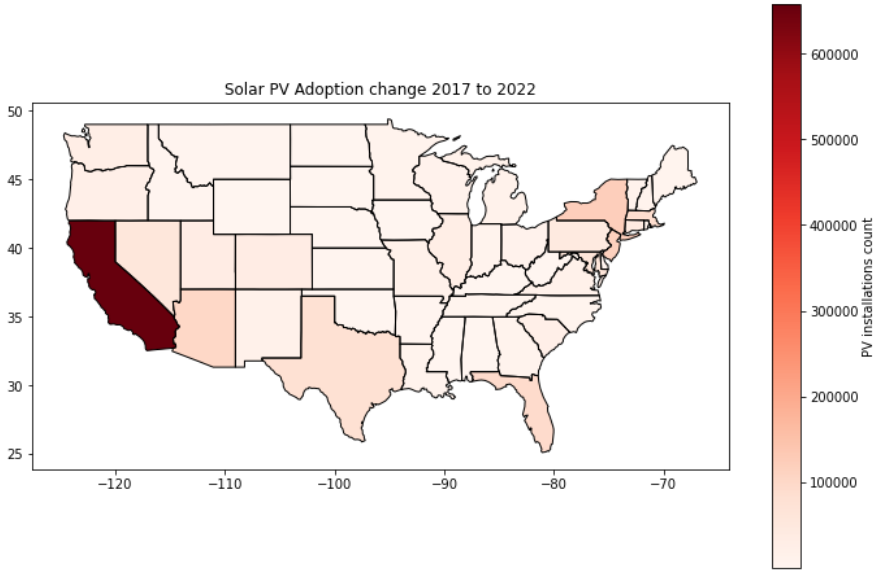

Figure 6.4: New Solar PVs (count) adopted across the country, state-wise. *Colors represent the amount of newly installed PV installations (2017-2022) in each state.*

Regarding the end-of-project goals: (1) we have constructed a set of GIS maps (called Solar Energy Atlas) containing information of both solar installations over time as well as distribution grids with high spatiotemporal granularity, together with the demonstration of incorporating different layers of maps for analyzing and modeling solar adoption and grid resilience; (2) we have analyzed the spatiotemporal pattern and underlying dynamics of solar adoption, with the effects of various solar incentives estimated at different income levels; (3) we have deployed the Solar Energy Atlas data to the Data Commons platform; (4) we have convened our first webinar on computer vision for power grids, and our second webinar on the application of the Solar Energy Atlas data.

8. Significant Accomplishments and Conclusions

Our significant accomplishments can be summarized as below:

- By developing new machine-learning-based solar panel identification algorithms, we have constructed the first and, so far, the most comprehensive solar PV spatiotemporal database covering the entire US. This is the first time that we obtained the exact GPS locations, size, subtype, and installation year information

for rooftop solar PVs across the U.S. By making this database publicly available, it is expected to serve as a valuable resource for solar PV growth forecasting, solar energy integration, and climate risk assessment of distributed energy systems at a spatially resolved scale.

- Utilizing the DeepSolar database, we have identified the non-linearity and heterogeneity in the dynamics of residential solar PV adoption. This has corrected the previous findings on solar adoption trends based on linear models and provided actionable insights for promoting solar adoption in an equitable way. Specifically, we find that, low-income communities not only started adoption later, but also more likely to get saturated at a lower adoption level. Performance-based incentives are effective in low-income communities while rebate programs are effective in high-income ones. This finding has provided important implications for tailoring solar energy incentives based on local income levels.

- We have developed a distribution grid GIS mapping algorithm using publicly available unstructured data as inputs, of which the effectiveness is verified both in the U.S. and Sub-Saharan Africa. This new algorithm can serve as a tool not only for guiding electricity access expansion in areas with limited electricity access (e.g., cities in Sub-Saharan Africa), but also for assessing the climate risk of grid infrastructures at a spatially resolved scale.

- Utilizing the distribution grid GIS mapping algorithm we have uncovered the non-uniform vulnerability of distribution grids to wildfires in California. In particular, we find that, at the same level of wildfire threats, low-income communities not only have less undergrounding protection of distribution lines, but also have higher vulnerability of the overhead part of grid infrastructure to wildfires and less distributed solar PV preparedness as the last-resort wildfire prevention approach. This has provided important policy implications for integrating socioeconomic status and climate-induced risks to make grid infrastructure adaptation approaches equitably affordable.

- We have incorporated the large-scale granular data on solar PVs and distribution grids produced by our algorithms into the Data Commons platform (https://datacommons.org) , enabling them to be accessible and user-friendly such that they can be easily correlated with other variables for engineering and socioeconomic applications.

Our impact is three-fold: (1) Models. Compared with conventional survey, crowdsourcing, or data reporting methods, our machine-learning-based data producing workflow to obtain granular data on solar energy and their associated infrastructure is automated, non-intrusive, and extensible to different countries. Our method relies on frequently-updated imagery and other open data hence the data produced in this project is easy to update. (2) Datasets. The large-scale, fine-grained datasets on solar installations and distribution grids are open, enabling our researchers, industry, and policymaking to develop various engineering or socioeconomic models and gain

insights. (3) Insights. The analyses we have conducted based on the produced data enabled the identification of effective solar incentives in low-income communities as well as equitable cost allocation schemes for improving power grid resilience.

9. Path Forward

We suggest future research in the following directions, based on the models, data, and insights gained from this project:

Future direction 1: A unified and automated framework for mapping and tracking DERs
In this project, we have developed algorithms that can be used for extracting granular spatiotemporal information about solar PVs from satellite images. Other types of DERs, such as EV chargers and battery storage could have different spatiotemporal adoption patterns, driving factors, or sensitivity to policies and incentives, but their information cannot be extracted from the same data source (i.e., satellite images). Instead, other types of unstructured yet widely available data such as building permits contain rich information about these DERs, but require different information extraction and mapping approaches. How to efficiently map and track different DERs from multiple types of data sources (images, text, and tabular data, etc.) to maintain a granular, up-to-date DER installation database could be of great interest to developers, utilities, and policy makers. The integration of advanced computer vision and natural language processing techniques can play an important role in developing such a unified and automated framework for mapping and tracking DERs.

Future direction 2: The patterns, driving factors, and policy effects for co-adoption of solar PVs and other DERs
The co-adoption of solar PVs and other types of DERs is becoming increasingly common, and has critical implications on grid operation and planning, financial incentive design, and energy justice. However, the co-adoption trends of these DERs have not yet been uncovered and studied at a large scale. Future direction 1 can help bridge the data gap. On top of this, an important research question that can be answered is: what are the underlying factors that shape the heterogeneity in the co-adoption rates of solar PVs and EV chargers (as well as solar PVs and battery storage) across places and time. This can further provide guidance for grid hosting capacity expansion, policy and incentive design, and other applications. The close collaboration of power system researchers, data scientists, and social scientists is essential to achieve this research goal.

Future direction 3: Investigating the effects of solar PVs and other DERs on climate resilience
The increasing adoption of solar PVs and other DERs can greatly reshape the power system resilience to climate-induced extremes (e.g., wildfires, hurricanes). Their effects can depend on a variety of factors (e.g., whether there is co-adoption of two or more types of DERs, electric demand) and can vary across different geographic locations and disaster types. Uncovering the effects of these DERs on climate resilience can be of great importance to ensuring reliable and equitable energy supply facing the increasing

threat of climate extremes yet requires highly granular geospatial and temporal information about DERs to perform correlational and even causal analyses. Therefore, we propose the development of a data-driven pipeline to extract such insights from large-scale data and provide user-friendly interfaces (e.g., API) for stakeholders (e.g., utilities) to get access to the needed information as well as actionable insights for policymakers to make informed decisions for reducing the vulnerability of communities to climate-induced extremes in an equitable way.

We also plan to keep engaging our industry advisory board members to identify new use cases of the technology we developed and potential needs for technology advancement.

10. Products

Publications:

1. Wang, Z., Arlt, M. L., Zanocco, C., Majumdar, A., & Rajagopal, R. (2022). Deepsolar++: understanding residential solar adoption trajectories with computer vision and technology diffusion models. *Joule*, 6(11), 2611-2625.
2. Wang, Z., Wara, M., Majumdar, A., & Rajagopal, R. (2023). Local and utility-wide cost allocations for a more equitable wildfire-resilient distribution grid. *Nature Energy*, 8(10), 1097-1108.
3. Wang, Z., Majumdar, A., & Rajagopal, R. (2023). Geospatial mapping of distribution grid with machine learning and publicly-accessible multi-modal data. *Nature Communications*, 14(1), 5006.
4. Wussow, M., Zanocco, C., Wang, Z., Prabha, R., Flora, J., Neumann, D., Majumdar, A., & Rajagopal, R. (2024). Exploring the potential of non-residential solar to tackle energy injustice. *Nature Energy*, 1-10.

Website:

1. DeepSolar: https://web.stanford.edu/group/deepsolar/home.html
2. Energy Atlas: https://web.stanford.edu/group/energyatlas/home.html

11. Project Team and Roles

PI:

Ram Rajagopal: overall management; student advising; idealization; manuscript revising; networking

Co-PI:

Arun Majumdar: student advising; idealization; manuscript revising; networking

Key Personnel:

Andrew Ng: student advising
Chin-Woo Tan: funding management; project management; regular reporting
June Flora: funding management; project management; regular reporting

Postdocs and Students:

Zhecheng Wang: idealization; project implementation; manuscript writing and revising; student mentoring; regular reporting
Chad Zanocco: project implementation; manuscript writing and revising; student mentoring; regular reporting
Rajanie Prabha: project implementation; manuscript writing and revising; regular reporting
Moritz Wussow: project implementation; manuscript writing and revising; regular reporting

Collaborators:
Ramanathan V. Guha: Provided support for Data Commons integration

12. References

1. Global Energy Observatory. http://globalenergyobservatory.org
2. Global Power Plant Database. http://datasets.wri.org/dataset/globalpowerplantdatabase
3. "Tracking the Sun" database. https://emp.lbl.gov/tracking-the-sun
4. Malof, J. M., Bradbury, K., Collins, L. M., & Newell, R. G. (2016). Automatic detection of solar photovoltaic arrays in high resolution aerial imagery. Applied energy, 183, 229-240.
5. Yuan, J., Yang, H. H. L., Omitaomu, O. A., & Bhaduri, B. L. (2016, December). Large-scale solar panel mapping from aerial images using deep convolutional networks. In 2016 IEEE International Conference on Big Data (Big Data) (pp. 2703-2708). IEEE.
6. Malof, J. M., Bradbury, K., Collins, L. M., Newell, R. G., Serrano, A., Wu, H., & Keene, S. (2016, November). Image features for pixel-wise detection of solar photovoltaic arrays in aerial imagery using a random forest classifier. In 2016 IEEE International Conference on Renewable Energy Research and Applications (ICRERA) (pp. 799-803). IEEE.
7. Yu, J., Wang, Z., Majumdar, A., & Rajagopal, R. (2018). DeepSolar: A machine learning framework to efficiently construct a solar deployment database in the United States. Joule, 2(12), 2605-2617.
8. Gurara, D., Klyuev, V., Mwase, N., & Presbitero, A. F. (2018). Trends and challenges in infrastructure investment in developing countries. International Development Policy| Revue internationale de politique de développement, (10.1).
9. Arderne, C., Zorn, C., Nicolas, C., & Koks, E. E. (2020). Predictive mapping of the global power system using open data. Scientific data, 7(1), 19.
10. Haklay, M., & Weber, P. (2008). Openstreetmap: User-generated street maps. IEEE Pervasive computing, 7(4), 12-18.
11. Liao, Y., Weng, Y., Liu, G., & Rajagopal, R. (2018). Urban MV and LV distribution grid topology estimation via group lasso. IEEE Transactions on Power Systems, 34(1), 12-27.
12. Deka, D., Backhaus, S., & Chertkov, M. (2016, June). Estimating distribution grid topologies: A graphical learning based approach. In 2016 Power Systems Computation Conference (PSCC) (pp. 1-7). IEEE.

13. Deka, D., Backhaus, S., & Chertkov, M. (2017). Structure learning in power distribution networks. IEEE Transactions on Control of Network Systems, 5(3), 1061-1074.
14. Weng, Y., Liao, Y., & Rajagopal, R. (2016). Distributed energy resources topology identification via graphical modeling. IEEE Transactions on Power Systems, 32(4), 2682-2694.
15. Yu, J., Weng, Y., & Rajagopal, R. (2017). PaToPa: A data-driven parameter and topology joint estimation framework in distribution grids. IEEE Transactions on Power Systems, 33(4), 4335-4347.
16. Yu, J., Weng, Y., & Rajagopal, R. (2018). PaToPaEM: A data-driven parameter and topology joint estimation framework for time-varying system in distribution grids. IEEE Transactions on Power Systems, 34(3), 1682-1692.
17. Scully, P. Smart Meter Market 2019: Global penetration reached 14%-North America, Europe ahead. https://iot-analytics.com/smartmeter-market-2019-global-penetration-reached-14-percent (2019).
18. Schmidt, E. H., Bhaduri, B. L., Nagle, N., & Ralston, B. A. (2018). Supervised classification of electric power transmission line nominal voltage from high-resolution aerial imagery. GIScience & Remote Sensing, 55(6), 860-879.
19. Gomes, M., Silva, J., Gonçalves, D., Zamboni, P., Perez, J., Batista, E., ... & Gonçalves, W. (2020). Mapping utility poles in aerial orthoimages using atss deep learning method. Sensors, 20(21), 6070.
20. Huang, B., Yang, J., Streltsov, A., Bradbury, K., Collins, L. M., & Malof, J. M. (2021). GridTracer: Automatic mapping of power grids using deep learning and overhead imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15, 4956-4970.
21. Zhang, W., Witharana, C., Li, W., Zhang, C., Li, X., & Parent, J. (2018). Using deep learning to identify utility poles with crossarms and estimate their locations from google street view images. Sensors, 18(8), 2484.
22. Krylov, V. A., Kenny, E., & Dahyot, R. (2018). Automatic discovery and geotagging of objects from street view imagery. Remote Sensing, 10(5), 661.
23. Kim, J., Kamari, M., Lee, S., & Ham, Y. (2021). Large-scale visual data–driven probabilistic risk assessment of utility poles regarding the vulnerability of power distribution infrastructure systems. Journal of construction engineering and management, 147(10), 04021121.
24. Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., & Yan, J. (2019). Siamrpn++: Evolution of siamese visual tracking with very deep networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4282-4291).
25. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
26. Integration Capacity Analysis Map. https://www.pge.com/en/about/doing-business-with-pge/interconnections/distributed-resource-planning-data-and-maps.html
27. Africa - Electricity Transmission And Distribution Grid Map. https://datacatalog.worldbank.org/search/dataset/0040465
28. Bass, F. M. (2004). The bass model. Manag. Sci., 50(12 Supplement), 1833-1840.

29. Antonio Olmos and Priyalatha Govindasamy. A practical guide for using propensity score weighting in r. Practical Assessment, Research, and Evaluation, 20(1):13, 2015.
30. Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113(523):1228–1242, 2018.
31. California Forest Observatory (2020). A Statewide Tree-Level Forest Monitoring System. Salo Sciences, Inc. San Francisco, CA. https://forestobservatory.com
32. Mann, M. L., Batllori, E., Moritz, M. A., Waller, E. K., Berck, P., Flint, A. L., ... & Dolfi, E. (2016). Incorporating anthropogenic influences into fire probability models: Effects of human activity and climate change on fire activity in California. PLoS One, 11(4), e0153589.
33. CAL FIRE. Fire probability for carbon accounting. *https://frap.fire.ca.gov/frap-projects/fire-probability-for-carbon-accounting*