JOUL, Volume *2*

Supplemental Information

DeepSolar: A Machine Learning Framework to Efficiently Construct a Solar Deployment Database in the United States

Jiafan Yu, Zhecheng Wang, Arun Majumdar, and Ram Rajagopal

1 SUPPLEMENTAL DATA

1.1 Satellite Imagery Dataset

1.1.1 Satellite Imagery

We use publicly available Google Static Maps API as our imagery source. Compared to the United States Geological Survey (USGS) orthoimagery that has resolution less than 30 cm, and has been used in previous works[6,7], Google Static Maps API provides satellite imagery with approximately 15 cm resolution covering the whole contiguous U.S. Such resolution is at the same scale of a single solar wafer and single solar panel usually contains dozens or more of such wafers. Another advantage of Google Static Maps API is that it is updated annually, enabling us to frequently update our results and track the growth of solar panel installations. We retrieve image tiles from Google Static Maps API as 320×320 pixels at zoom 21. The side length of an image tile varies from 15 m to 22 m from north to south. Though the nominal resolution of the downloaded image tile is smaller than 15 cm, the actual resolution of the acquired satellite images is still 15 cm and the downloaded image tile is extrapolated from the original satellite imagery with 15 cm resolution.

1.1.2 Constructing Process

The dataset is constructed by integrating both CNN automatic identification and manual annotation. Initially, we randomly collect 320,000 satellite images in over 50 different cities/towns with Google Static Maps API and manually labeled them with Amazon Mechanical Turk (AMT), a crowdsourcing platform for solving labor-intensive task. However, one challenge is that only less than 1% of the samples are positive, which is due to the fact that solar panels are very sparse compared to the total area of a region. We solve this obstacle by training a preliminary classification model based on VGG-1637 network with all positive samples and a small fraction of the negative samples in this dataset and using it to identify solar panels around San Francisco Bay Area. 55,193 image tiles are identified to be positive with this model. Leveraging AMT, false positive samples are removed from the positive-predicted samples. Finally, we construct a dataset containing 472,953 samples, of which 50,507 are actually positive.

Although instructions are provided for AMT users to identify solar panels from satellite images, invariably there are mistakes. To increase the reliability of manual labeling, we use AMT following the scheme as shown in Fig. S1. All images are identified by two qualified users. The results from the users who fail the accuracy test are not used. Extra details, such as the building information, are examined for identifying the samples that two users disagree on.

1.1.3 Dataset Statistics

Our dataset consists of a training set (366,467 samples), a validation set (12,986 samples) and a test set (93,500 samples). The basic statistics of the dataset are shown in Table S1. Images in test set are randomly sampled in 35 regions across the contiguous U.S., and they are from the regions totally different from those in training set, even though both of them may be in the same city/town. To evaluate the ability of area estimation, each test sample is also annotated with ground truth regions of solar panels besides image-level label.

1.2 Other Data

We collect additional data and integrate to obtain census-tract-level environmental and socioeconomic factors to support estimating solar deployment density. We utilize multiple public data sources to create the following factors at tract level:

1.2.1 Indicators Reflecting the Cumulative Effects of Statewide Incentives on Residential Solar Deployment

We collect data of the number of years since the start of different types of statewide energy policies/incentives that aim at boosting solar deployment. These indicators reflect the cumulative effects of incentives and are used in previous works[15,16,38]. They are: number of years since the start of net metering, number of years since the start of feed-in tariff, number of years since the start of rebate program, number of years since the start of property tax incentives, number of years since the start of sales tax incentives, number of years since the start of corporate tax credit programs. If a type of incentive never exists in a state, then the corresponding indicator is set to be zero. Data is obtained from *dsireusa.org*.

1.2.2 Average Residential Electricity Retail Rates/Prices over the Past Five Years

Average residential retail electricity rate/price data over the past five years is obtained from U.S. Energy Information Administration (EIA)[39].

1.2.3 Demographic Factors

Demographic factors, such as income, education, population density, are collected from American Community Surveys (2011-2015)[40]. More specifically, we use the following demographic variables: population (T001_001), area (T003_001), age (T007_001 to T007_013), household type (T017_001, T017_002, T017_007), average household income (T059_001), poverty status (T113_001, T113_001), Gini index (T157_001), employment rate (T037_001), educational attainment (T025_001 to T025_008), dropout rate (T030_001 to T030_003), race (T013_001 to T013_008), occupation(T049_001 to T049_014), house heating fuel (T099_001 to T099_008), house occupancy status (T094_002, T094_003, T095_001, T095_003), average household size (T021_001), median housing unit value (T101_001), median housing unit gross rent (T104_001), mortgage status (T108_001, T108_002, T108_008), means of transportation to work (T128_001 to T128_010), travel time to work (T148_001 to T148_008), health insurance status (T145_001 to T145_005).

Specifically, we extract average household size, median housing unit gross rent, median housing unit value, average household income, years of education, Gini index, population density, poverty rate, employment rate, ratio of Asian, ratio of black/Africa American, ratio of white rate, ratio of American Indian/Alaska native, ratio of native Hawaiian/other Pacific islander, ratio of two or more races, racial diversity, ratio of using coal/coke/wood as heating fuel, ratio of using electricity as heating fuel, ratio of using solar energy as heating fuel, ratio of using fuel/oil/kerosene as heating fuel, ratio of using gas rate, ratio of using no heating fuel, less than high school rate, high school graduate rate, college rate, bachelor rate, master rate, professional school rate, doctoral rate, median age, ratio of age between 5 and 9, ratio of age between 10 and 14, ratio of age between 15 and 17, ratio of age between 18 and 24, ratio of age between 25 and 34, ratio of age between 35 and 44, ratio of age between 45 and 54, ratio of age between 55 and 64, ratio of age between 65 and 74, ratio of age between 75 and 84, ratio of age more than and 85, ratio of household with family, in-school rate for 16 to 19 year-old, ratio of construction-related occupation, ratio of public-related occupation, ratio of information-related rate, ratio of finance-related occupation, ratio of education-related occupation, ratio of administrative-related occupation, ratio of

manufacturing-related occupation, ratio of wholesale-related occupation, ratio of retail-related occupation, ratio of transportation-related occupation, ratio of arts-related occupation, ratio of agriculture-related occupation, ratio of vacant housing unit, ratio of owner-occupied housing unit, ratio of mortgage, ratio of working at home, ratio of using car, ratio of walk to work, ratio of using carpool, ratio of using motorcycle, ratio of using bicycle, ratio of using public transportation, ratio of less than 10 min to work, ratio of 10-19 min to work, ratio of 20-29 min to work, ratio of 30-39 min to work, ratio of 40-59 min to work, ratio of 60-89 min to work, ratio of taking public health insurance, ratio of taking no health insurance, average travel time to work.

In addition to these metrics, we evaluate the education level of a census tract by computing the average years of education. In the United States, completing middle school usually takes 8 years; the length of completing a typical high school, community college, undergraduate university, masters, Ph.D. usually takes 4, 2, 4, 2, 5 years, correspondingly. By summing the numbers together, we can compute the number of years in school for a person with different highest degrees, which is shown in Table S2. For each census tract, we can compute the average years of education by:

$$\text{Average years of education} = \frac{1}{N}\sum_{i=1}^{N}\text{years of education for individual i} \quad (1)$$

where N is the number of population for the census tract. By grouping individuals by different highest obtained degrees, average years of education can be calculated as a weighted sum:

$$\begin{aligned}\text{Average years of education} &= \frac{1}{N}\sum_{i=1}^{M} N_j \times \text{years of education for degree type } j \\ &= \sum_{i=1}^{M}\frac{N_j}{N} \times \text{years of education for degree type } j \quad (2)\end{aligned}$$

where N_j is the number of population with degree type j as their highest degree. The ratio N_j/N can be found from American Community Surveys. Combining the ratios and the years of education for each type of degree, we can compute the average years of education for each census tract as a scalar proxy of education level:

$$\begin{aligned}\text{Average years of education} = {} & \text{less than high school rate} \times 8 + \text{high school rate} \times 12 \\ & + \text{college rate} \times 14 + \text{bechalor rate} \times 16 + \text{master rate} \times 18 \\ & + \text{professional school rate} \times 21 + \text{doctoral rate} \times 21 \quad (3)\end{aligned}$$

We also use racial diversity (Simpson's Diversity Index) as a demographic factor by calculating:

$$\text{Racial diversity} = \sum_{i=1}^{N} r_i(1 - r_i) \quad (4)$$

where r_i is the ratio of some race in the total population.

1.2.4 Political attitude

Political attitude is represented by DEM voting percentage, and GOP voting percentage in 2016 election. They are obtained from theguardian.com[41] and townhall.com[42].

1.2.5 Environmental condition

Environmental data comes from the NASA Surface Meteorology and Solar Energy[43] with spatial resolution of 1˚. We extract various factors including air temperature, relative humidity, solar radiation, atmospheric pressure, wind speed, earth temperature, elevation, earth temperature amplitude, number of frost days, heating degree day, cooling degree day.

2 SUPPLEMENTAL EXPERIMENTAL PROCEDURES

2.1 More details on using image classification for system detection

To tackle the highly imbalanced class distribution in the dataset, we utilize cost-sensitive learning in the image classification. Cost-sensitive learning gives different penalties to different misclassifications in constructing the loss function. In our work, we simply give more penalty to the misclassifications of positive samples by formulating the loss function:

$$L = \sum_i [\alpha \mathbf{1}(y_i = 1)\mathrm{softmax}(y_i, f(x_i)) + \mathbf{1}(y_i = 0)\mathrm{softmax}(y_i, f(x_i))] + \text{regularization term} \quad (5)$$

where y_i is the ground truth label and x_i is the input image. f represents the mapping of CNN from input image to an output confidence score. $\mathbf{1}$ is the indicator function. Softmax represents the softmax loss of a sample. Lower confidence score of true class results in higher softmax loss. α is the penalty coefficient given to the misclassifications of positive samples, which is larger than 1.

For training the classifier, each image is resized to 299×299 pixels before being fed into the CNN. Each training sample is rotated with an angle randomly selected among 0°, 90°, 180°and 270° for data augmentation. For training the basic classification framework (Inception-v3), we use RMSProp optimizer with decay of 0.9, momentum of 0.9 and epsilon of 0.1. The initial learning rate is 0.001 and decays every 5 epochs with a factor of 0.5. The batch size is 32. The penalty coefficient α we use is 10. Training is stopped after 185,000 steps.

2.2 More details on semi-supervised size estimation

2.2.1 Drawbacks of fully-supervised segmentation

Traditional fully-supervised approaches require ground truth annotations of object regions for training the segmentation capability[44,45]. However, manually annotating large number of object regions in the training set is very time-consuming and human-force intensive. Furthermore, fully-supervised segmentation involves the convolution-deconvolution process[46], substantially increasing the computational time. These drawbacks make the traditional fully-supervised approaches impractical for nationwide solar panel identification.

2.2.2 Class Activation Map

Previous work[35] shows that semi-supervised object localization and segmentation can be achieved by generating Class Activation Map (CAM) based on classification model. In CNN, the output of each convolutional layer is a stack of feature maps that represents activation of different features. We denote the shape of feature

maps as $h \times w \times n$. h is the height, w is the width of each feature map and n is the number of feature maps in a stack. For classification, we perform a global average pooling (GAP) firstly by simply taking the average value of each feature map, resulting in a $1 \times 1 \times n$ vector H, then we use a linear classifier mapping this vector to the classification scores S:

$$H^{1\times n}\ W^{n\times 2} = S^{1\times 2} \tag{6}$$

Bias terms are ignored in this linear classifier. CAM can be generated with the weighted sum of the feature maps, and the weights are just the entries in W connected to the positive class, as is illustrated in Fig. S2. Intuitively, CAM can be regarded as the linear combination of the visual patterns indicating the solar panel.

2.2.3 Greedy layer-wise training

One problem with the method discussed above is the fact that only the salient part of solar panel, which is mostly activated, can be extracted. This results in underestimation of the solar panel size. We improve the method by utilizing one important property of CNN: features learned at low-level hierarchy (upstream layers) are low-level and general, like edges or basic shapes, while features learned at high-level hierarchy (downstream layers) are high-level and specific. Therefore, feature maps at low-level hierarchy are more complete but noisy, while feature maps at high-level hierarchy are more discriminative but incomplete. This trade-off is illustrated in Fig. S3.

We can break the trade-off by greedily extracting features at low-level hierarchy to generate a both complete and discriminative CAM for segmentation. Specifically, we train a single "convolutional layer-GAP-linear classifier" structure for image-level classification at a time, based on a pre-trained network for classification. Then we discard the GAP and linear classifier and add a new "convolutional layer-GAP-linear classifier" structure at the end of the last convolutional layer, and train the newly added layers separately. When training a single "convolutional layer-GAP-linear classifier" structure, we keep all other layers completely fixed, thus weights and biases of those layers are not updated. The basic Inception-v3 architecture and greedy layer-wise training process are illustrated in Fig. S4.

The benefit of greedy layer-wise training is shown in Fig. S5. We can find that greedy layer- wise training can significantly improve the quality of the CAMs. The regions of solar panel are more complete, the resolution is higher, and the boundaries of are much clearer, which contributes to more accurate area estimation of solar panels. More examples of segmentation and area estimation is shown in Fig. S6. Note that another benefit of our model is that both classification and segmentation results can be output in a single forward pass, and the segmentation branch will be executed only if the input image is classified to be positive. Therefore, our algorithm significantly reduces the computing time.

For training the segmentation branch, we used a Gradient Descent optimizer with a constant learning rate of 0.005. The batch size is 64. First "convolutional layer-GAP-linear classifier" structure is trained for 20,000 steps and second "convolutional layer-GAP-linear classifier" structure is trained for 15,000 steps. The threshold we use for segmentation is 0.37.

2.2.4 Merging to concatenate large solar systems

As a large-scale solar panel can be split to several image tiles, we use a simple recursive algorithm to find adjacent pieces of solar panels and merge them to form a solar system, as is shown in Algorithm 1 (at the end of SI).

2.3 Performance evaluation of classification and segmentation

Both classification and size estimation performances of our model are evaluated on a test set containing 93,500 randomly-sampled images across the U.S. Precision and recall are used as metrics to measure the classification capability. They are defined as:

$$\text{Precision} = \frac{\#\,\text{True Positive}}{\#\,\text{True Positive} + \#\,\text{False Positive}} \tag{7}$$

$$\text{Recall} = \frac{\#\,\text{True Positive}}{\#\,\text{True Positive} + \#\,\text{False Negative}} \tag{8}$$

"True Positive" means correctly predicted positive samples, "False Positive" means positive (wrong) prediction on negative samples, and "False Negative" means negative (wrong) prediction on positive samples. Precision measures the ratio of correct predictions among positive predictions. Recall measures the ratio of actual positive samples that can be identified. Since the output of the classifier is a predicted probability of being positive for an image, by adjusting the threshold probability, precision-recall curves can be generated (Fig. S7). Precision and recall are trade- off, and they can both reach around 90% by setting the probability threshold to be 0.5. With this threshold, the precision is 93.1% in residential areas and 93.7% in non-residential areas. The recall is 88.5% in residential areas and 90.5% in non-residential areas. Therefore, we use the 0.5 as the threshold probability for solar panel identification. The raw confusion matrices for this threshold are shown in Table S3 (residential areas) and Table S4 (non-residential areas).

Besides image-level metrics, we also evaluate the classification performance at region level. By counting the estimated numbers of tiles containing solar at region level and comparing them with the true number of solar tiles, we find that the region-level relative counting error is within 1% for non-residential areas and within 5% for residential areas (Fig. S8), suggesting that the model has smaller error rate at region level than image level.

For size estimation, the image-level estimation residual plots are shown in Fig. S9. The mean residual is -0.949 m^2 for residential areas and -1.919 m^2. Compared to the 10 to 100 m^2 level of solar panel area in an image tile, such image-level area estimation performance can be regarded as nearly unbiased. At region level, difference between estimated and ground truth total solar panel area for different test regions are shown in Fig. S10. In most regions, the relative area estimation error rate is within 5%. To further quantify the area estimation performance, we introduce mean relative error (MRE), which is defined as:

$$\begin{aligned}\text{MRE} &= \frac{\sum_i^{\#\text{ true positive}} (\text{true area}_i - \text{estimated area}_i)}{\sum_i^{\#\text{ true positive}} \text{true area}_i} \\ &= \sum_i^{\#\text{ true positive}} \frac{\text{true area}_i}{\sum_i^{\#\text{ true positive}} \text{true area}_i} \frac{\text{true area}_i - \text{estimated area}_i}{\text{true area}_i}\end{aligned} \tag{9}$$

As is shown, MRE is equivalent to the weighted average of relative area estimation error. The reason we use weighted average instead of average is to balance the difference of true area among different test regions. The MRE for solar panel area estimation at region level is 3.0% for residential areas and 2.1% for non-residential areas.

2.4 Nationwide solar panel identification

2.4.1 Database coverage

We feed the DeepSolar framework with 1.1 billion image tiles covering all U.S. urban areas defined by U.S. Census Bureau[47] and all areas with nightlight intensity greater than 128 out of 255. U.S. urban areas contain populated census tracts and their adjacent territory, accounting for 80.7% of the total population[47]. On the other hand, nightlight intensity is a good proxy for building and population density[48]. We utilize the nightlight map released by NASA in 2016[49] that measures nightlight intensities in a range of 0 to 255. For non-urban areas, currently we do not include regions with nightlight intensity lower than 128 to provide an initial database. To estimate the proportion of solar installations covered in the detected range, we calculate the ratio of solar panel image tiles located in urban areas γ for each nightlight intensity between 128 and 255. γ is defined as:

$$\gamma = \frac{\#\ \text{image tiles in urban areas containing solar panels}}{\#\ \text{image tiles containing solar panels}} \tag{10}$$

We then fit a linear regression between γ and nightlight intensities (128-255) and use it to estimate γ in lower-nightlight-intensity (0-128) regions. The fitting result is as follow:

$$\gamma = 0.0136 \times \text{nightlight intensity} + 0.597 \quad (R^2 = 0.67) \tag{11}$$

Based on the ratio of solar panel image tiles located in urban areas, we can estimate the total number of solar panel image tiles in lower-nightlight-intensity (0-128) regions:

$$\text{Estimated \# image tiles containing solar} = \frac{\#\ \text{image tiles in urban areas containing solar panels}}{\gamma} \tag{12}$$

Then we can estimate the proportion of image tiles containing solar panels covered in the detected range. Result shows that we can cover over 95% image tiles containing solar panels in the contiguous U.S. (Fig. S11). After the submission of the article, we will continue to use the model to scan the lower-nightlight-intensity (less than 128) regions and thus increase the solar installation records in our database.

2.4.2 Estimation of Total Number of Solar

2,227,322 image tiles are identified to contain solar in the contiguous U.S. Here we calculate the expected value of the actual total number of positive image tiles. We regard each image tile as an independent random variable with Bernoulli distribution. For positive-predicted sample, its probability of being the actual positive sample is $p_{11} = p(y = 1 \mid \hat{y} = 1)$, where y denotes actual class and $\hat{y}$ denotes predicted class. For negative-predicted sample, its probability of being the actual positive sample is $p_{10} = p(y = 1 \mid \hat{y} = 0)$. Thus the total number of positive image tiles is a random variable with binomial distribution. The total number of image tiles is n, among which n_1 tiles are positive-predicted and n_0 tiles are negative-predicted.

We know n=1,091,548,472, n_1=2,227,322, precision=0.872, recall=0.890. Then we have:

$$n_0 = n - n_1 \tag{13}$$

$$TP = n_1 \times precision \tag{14}$$

$$FP = n_1 - TP \tag{15}$$

$$FN = \frac{TP}{recall} - TP \tag{16}$$

$$TN = n_0 - FN \tag{17}$$

$$p_{11} = p(y = 1 \mid \hat{y} = 1) = \frac{TP}{TP + FP} \tag{18}$$

$$p_{10} = p(y = 1 \mid \hat{y} = 0) = \frac{FN}{FN + TN} \tag{19}$$

With p_{11} and p_{10}, we can calculate the expected value of actual total number of positive image tiles m and its standard deviation **σ**:

$$m = n_1 p_{11} + n_0 p_{10} \tag{20}$$

$$\sigma = \sqrt{n_1 p_{11}(1 - p_{11}) + n_0 p_{10}(1 - p_{10})} \tag{21}$$

Finally, the expected value of actual total number of positive image tiles m is 2,197,164 and its standard deviation **σ** is 692.

2.5 Predictive models for solar deployment density at the census tract level

2.5.1 Literature on solar deployment

Classical solar deployment analysis utilizes field surveys. Recent survey-based solar deployment researches include surveys in Texas, United States with 365 responses[17]; in Ontario, Canada with 298 responses[18]; in Wisconsin, United States with 36 responses[50]; in Austria with 52 responses[51]; in Netherlands with 817 responses[19]; in Sweden with 14 responses[20]; in California, United States with 380 responses[21]; in Arizona, California, New Jersey, New York, United States with 904 responses[22]; and in Austria and Italy with 22 responses[23]. The identified key correlated factors with high solar deployment density include higher education level[17], higher income[17], belief of financially prudency[17], desire to help the environment[17,22,23], cost of installation[18,19], solar radiation[18], enjoyment of technical innovation[22,50], and knowledge about solar[19]. However, survey-based approaches typically contain at most thousands of data samples, limiting the capability of generalization and building quantitative models. Most of the results from surveys are only qualitative and descriptive. Furthermore, the models are not built to achieve good out-of-sample predictions.

More recently, residential solar installation data from incentive programs, utilities, and government has been used for analyzing solar deployment. Existing research use datasets from Open PV Project[15,16,26], California Solar Initiative (CSI)[24] and California Center for Sustainable Energy (CCSE)[25] in the United States, Deutsche Gesellschaft für Solarenergie (DGS)[14] and Amprion[28] from Germany, and VERG from Belgium[27]. However, the existing datasets in the United States suffer from location resolution restricted to zip-code level or county level.

Existing papers build parametric models, in particular, linear and log-linear models to capture the correlation between socioeconomic factors and solar deployment density. These models usually provide a single parameter for each factor and with

qualitative and even only directional (positive or negative) conclusions. Solar radiation[14–16,24,26], electricity rates[15,16], number of incentives[15,16,26], average home value[15,16,27], owner-occupied home rate[15,16,24–26], income[15,27,28], education level[15,16,26,28], and democrat voters ratio[15,16] are observed to be positively correlated with solar installation. Income inequality[27] is also observed to be positively correlated with solar installation. Housing density is observed to be negatively correlated[15,16] or positively correlated[14].

However, the explained variance by these models is low. The reported in-sample R2 are for example: 0.33[16,25], 0.55[24], 0.57[28], and 0.59[26]. The low accuracy of these models hinders their use in predictive analysis. In the main text and SI 3.2 we observe that solar deployment density is nonlinearly related to various socioeconomic and environmental variables. Moreover, factors interact. For example, deployment trends for income are different when conditioned on solar radiation. Linear or log-linear models that utilize individual parameters without interactions can lead to poor fits and potentially incorrect conclusions. Existing models are also limited to utilizing zip-code or county level data so the number of regression data points are usually less than a few thousand. Variability within zip-code or county is missed, resulting in poor granularity. In addition, the number of solar installations used for analysis in the United States is limited, for example to 9,000[25], 58,841[15] or 118,471[24]. Another barrier to developing models is that most of the datasets do not provide the information whether a solar installation is residential or not. In many cases system size is not provided, even if provided, a 10kW capacity cut-off is typically used to classify installations as either residential or commercial resulting in potential inaccuracies.

2.5.2 More details on SolarForest and SolarNN

The main text describes SolarForest, a machine learning model to predict census tract level solar deployment density utilizing socioeconomic and environmental factors as inputs. SolarForest is a two-stage model integrating a Random Forest classifier and a Random Forest regressor. The prediction process is shown in Algorithm 2 (at the end of SI), where X denotes the vector of census tract level predictors, c denotes the classification result, and y denotes the final result (solar deployment density). The classifier contains 100 decision trees, and the regressor contains 200 decision trees.

SolarNN is another high-accuracy machine learning model to predict solar deployment density with socioeconomic and environmental factors. SolarNN is a feed forward neural network consists of an input layer with dimension 94, five fully connected hidden layers with 94 neurons in each layer, and a final layer with scalar output. The out-of-sample cross-validation R^2 of SolarNN is 0.722. Fig. S12 shows the feature importance ranking of SolarNN. Compared with feature importance rankings of SolarForest shown in Fig.6b and Fig.6c in the main text, we find that population density, daily solar radiation, and relative humidity are among the most important features for both of these two predictive models.

SolarForest and SolarNN both achieve a R^2 higher than 0.720 in 10-fold cross validation. Such accurate predictive models can be further utilized to predict "what-if" scenarios, and thus provide a valuable reference for policymakers and solar companies to explore the potential of solar deployment of a region according to local environmental and socioeconomic factors. Also, as we will use DeepSolar to update solar deployment database annually, we can retrain SolarForest and SolarNN to predict the yearly increment of solar deployment in every census tract instead of the cumulative amount, making our machine-learning-based predictive

models even more powerful.

3 CLARIFICATIONS ON TREND ANALYSIS

3.1 Analyze demographic trends conditioning on solar radiation

Through univariate scatter plots on full dataset (Fig. S13), we find that range of solar deployment density (*y*-axis) along with solar radiation is at least 5 times larger than any other factors (average household income, population density, etc.). Therefore, we do correlation analysis of other demographic/socioeconomic factors conditioning on solar radiation to exclude its significant impact.

3.2 Demographic trends with solar deployment

We focus our investigation in the relationship between solar deployment density and four demo- graphic factors to illustrate the value of the DeepSolar database: average annual household income ($), average years of education, Gini index, and population density (capita/mile^2), conditional on solar radiation. Each trend is built by grouping tracts into solar radiation levels defined by Fig. 4 in the main text. In SI 2.5.1, we observed that current literature focuses on zip-code or county level analysis whereas with the DeepSolar dataset, tract level observations can be made, highlighting trends more clearly.

Solar deployment density increases with average household income across all solar radiation groups (Fig. S14a, Fig. S14b, Fig. S14c). The group that has high solar radiation on average can be benefited from high solar power generation potential, thus it tends to have the highest long- term returns from solar deployment and the solar system is likely to be long-term profitable, hence the deployment density should be rather independent of income. Yet, we still observe a strong increasing trend for this group indicating that perhaps upfront cost is still a major financial burden for solar deployment. Finally, in high-radiation regions, deployment saturation or even decrease is observed at high income levels (>$150,000) suggesting other limiting factors.

We observe solar deployment density increases with education level as well (Fig. S14d, Fig. S14e, Fig. S14f). The trend might be partially explained by the strong correlation between education level and household income. To decompose these two factors, we separate their relationships by additionally grouping tracts by income and observing the education-solar deployment correlation given a certain level of household income. Fig. S15 shows that higher education level does boost solar deployment, but mainly in the areas with low solar radiation and low- to medium- level average household income. Once residential solar systems are long-term profitable (i.e., solar radiation is rich), education level shows no positive correlation with deployment density anymore.

Solar deployment shows a negative trend with Gini index (Fig. S14g, Fig. S14h, Fig. S14i). However, in contrast with education, Gini index only correlates with solar deployment in regions with high solar radiation, if excluding the impact of income (Fig. S16). This finding indicates that income equality promotes solar deployment in regions where solar systems are long-term profitable.

The relationship between population density and solar deployment also shows very similar trends across different subgroups (Fig. S14j, Fig. S14k, Fig. S14l). The peak of the solar deployment occurs when the population density is approximately 1000 capita/mile^2. The increasing trend before 1000 capita/mile^2 could potentially be explained by the feasibility of solar service companies, and by the proven peer effects[52,53]. The decreasing trend after the peak density could be due to the lack of

availability of rooftops at high densities since we normalize solar deployment by number of households. We check this hypothesis utilizing rooftop count data obtained from Googles Sunroof Project for the available tracts (48,722 tracts). Rooftop count and household count are highly correlated (Fig. S18), thus number of households is an excellent proxy for number of rooftops. When solar deployment density is measured as number of systems per number of rooftops, the trend with respect to population density is similar and still decreasing when population density is over 1000 capita/mile^2 (Fig. S17). We also computed trends with respect to the remaining factors and they are similar as well. This confirms that solar deployment at most census tracts is far from saturation rates and that deployment density as measured by systems per household is an appropriate solar deployment metric.

We expect the development of our dataset to encourage the community to uncover more relationships between solar deployment and other socioeconomic factors including political attitude and racial composition of tracts.

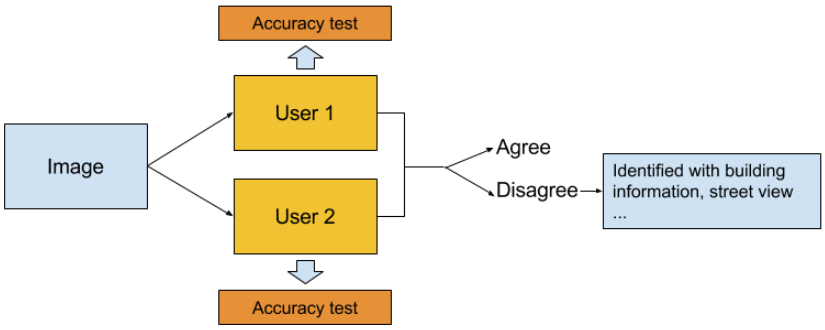

Figure S1: The framework of manual labeling with Amazon Mechanical Turk (AMT)
All images are identified by two users who pass the accuracy test. Samples with disagreement are identified with more comprehensive information.

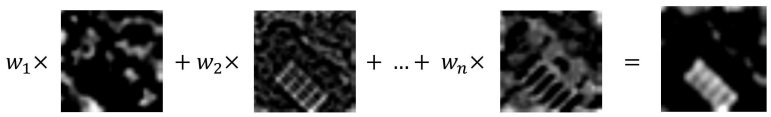

Figure S2: Class Activation Map (CAM)
The leftmost is the original input image. Feature maps are summed up with weights in linear classifier to generate CAM (rightmost), highlighting the discriminative region of solar panel.

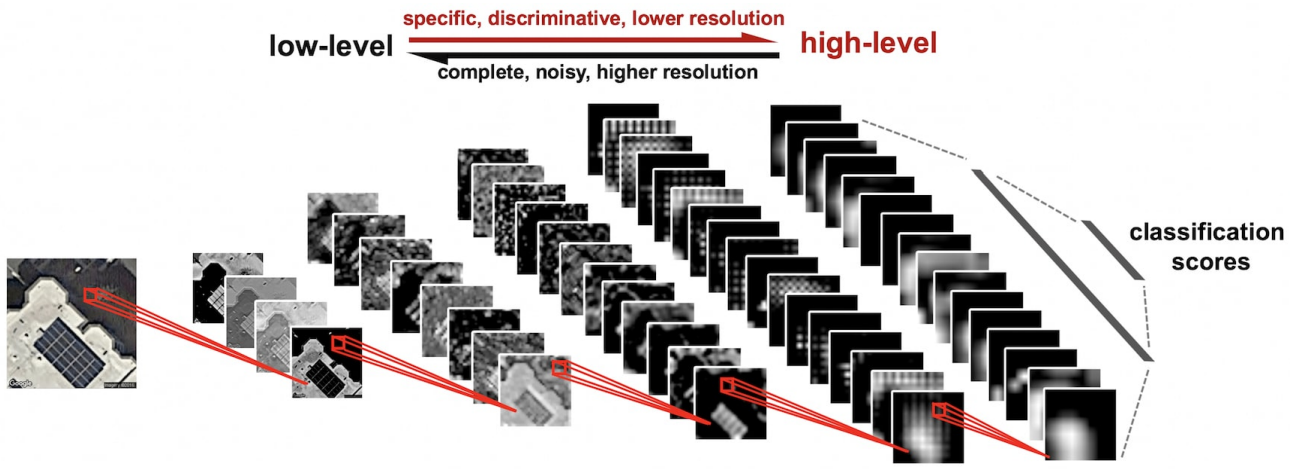

Figure S3: The feature hierarchy of CNN
From raw image input to class score output, different layers extract features of different levels. The feature maps at low-level hierarchy are complete, high-resolution but noisy, while the feature maps at high-level hierarchy are specific, discriminative but low-resolution.

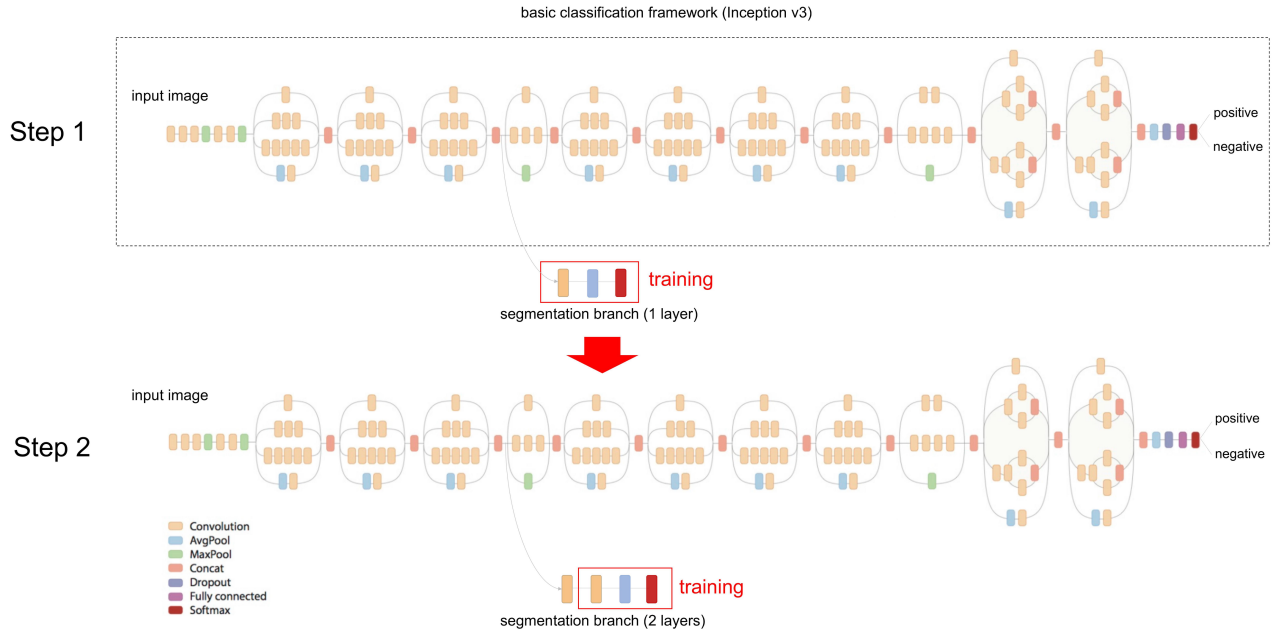

Figure S4: Inception-v3 framework and greedy layer-wise training

The layers in the dashed box form the basic classification framework (Inception-v3 model). Greedy layer-wise training is performed in two steps (labelled in the left in the figure). Step 1: Keep all layers fixed but train a single "convolutional layer-GAP-linear classifier" structure in the segmentation branch. Step 2: Add another "convolutional layer-GAP-linear classifier" structure at the end of the segmentation branch and train it with all other layers fixed.

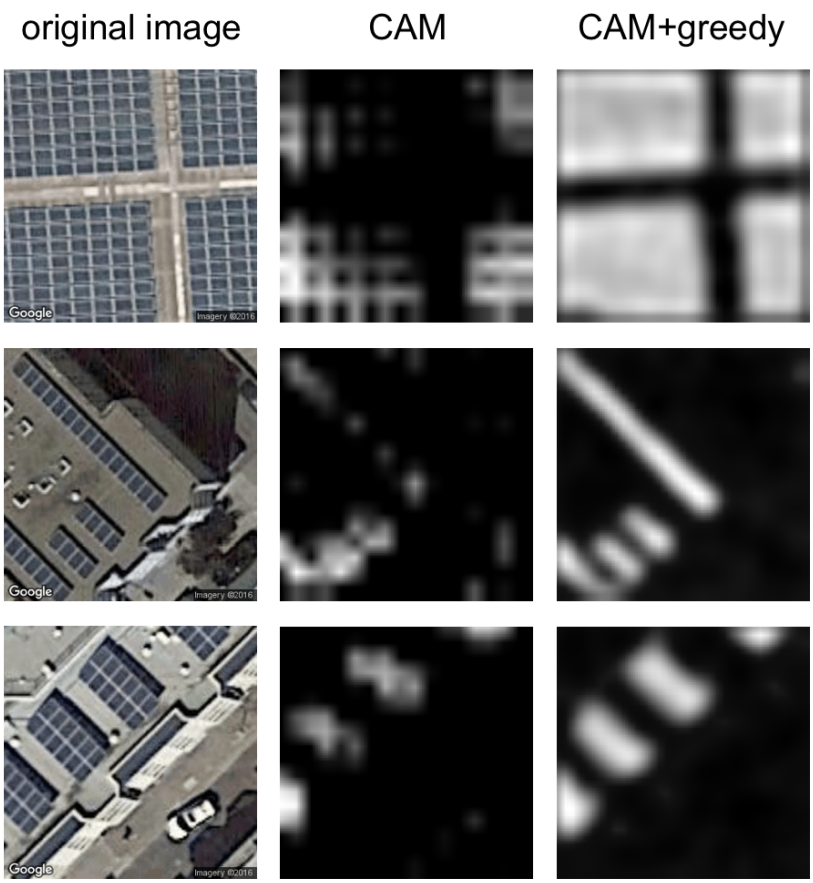

Figure S5: Benefit of greedy layer-wise training

The left column contains original images. The middle column are the Class Activation Maps (CAMs) of the original images without greedy layer-wise training. The right column are the CAMs of the original images with greedy layer-wise training.

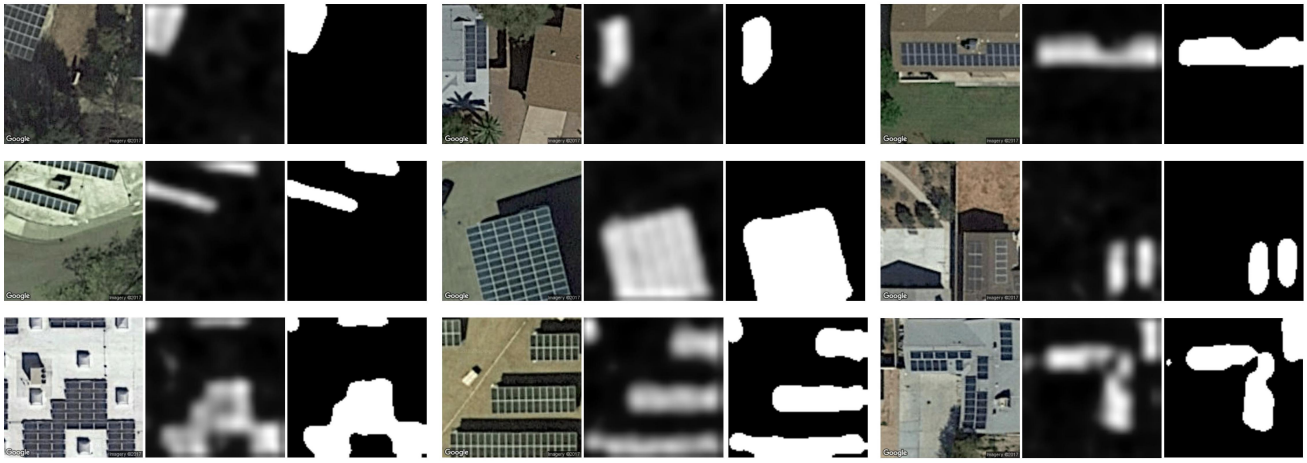

Figure S6: More examples of CAMs and the segmentation results in test set

For each group, the left one is the original satellite image retrieved from Google Static map API, the middle one is the Class Activation Map (CAM), and the right one is the segmentation result. A simple threshold is set to generate the segmentation result from the CAM.

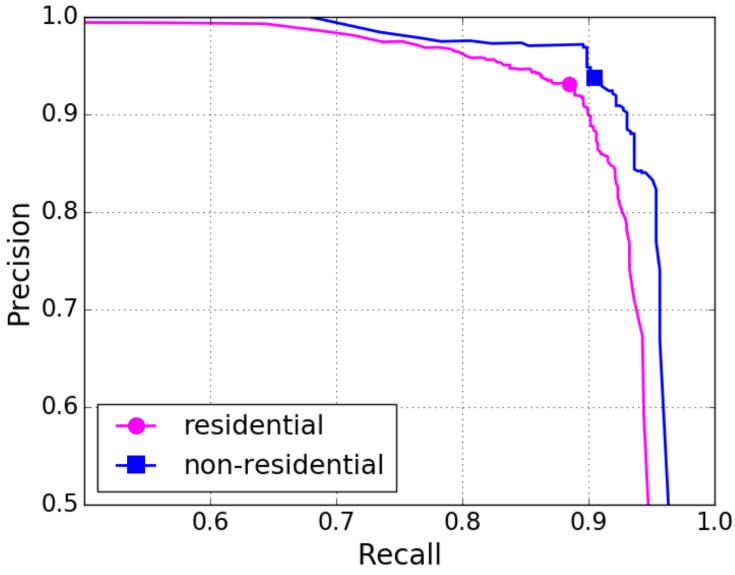

Figure S7: Precision-recall curve of the image classifier for solar panel identification in residential and non-residential areas

Setting the threshold probability to different values results in different precision and recall, generating the precision-recall curves. The square and circle denote the state we use with threshold probability of 0.5, which reaches around 90% recall and precision in both residential areas and non-residential areas.

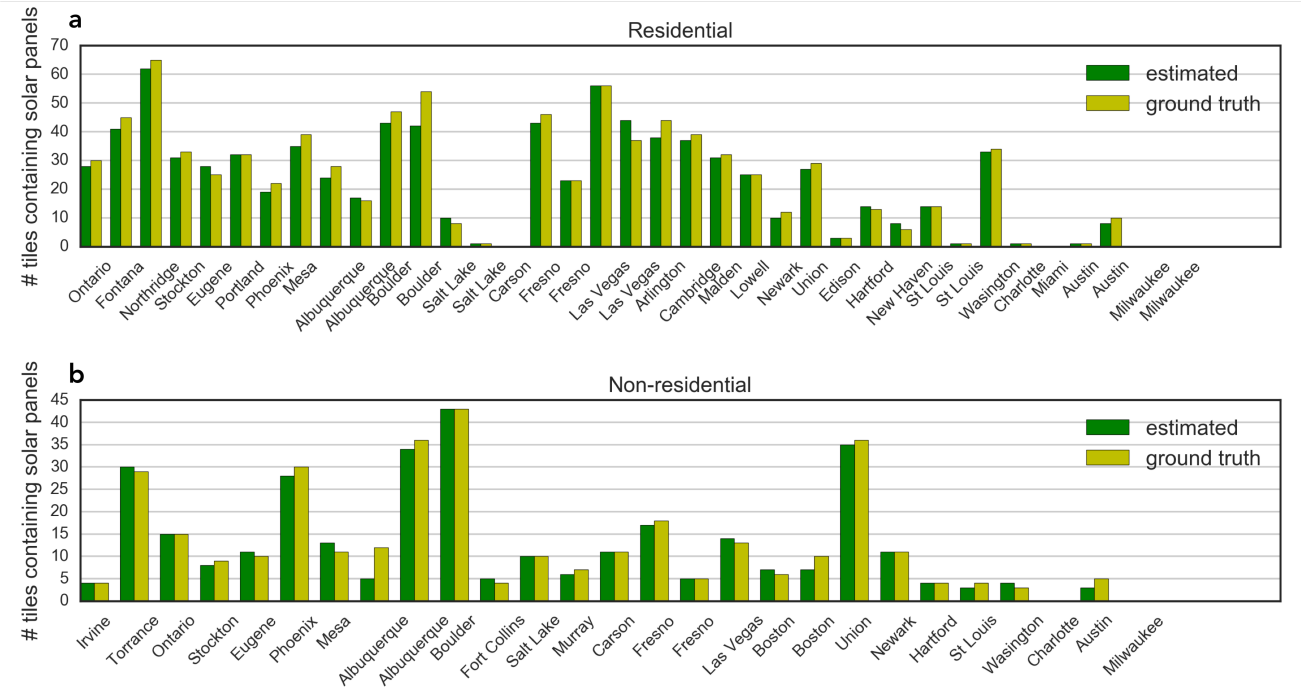

Figure S8: Region-level difference between estimated and actual number of image tiles containing solar panels

a. For residential areas. **b.** For non-residential areas. Some test regions have same name because they are in the same city but compiled from different districts (collection of census tracts).

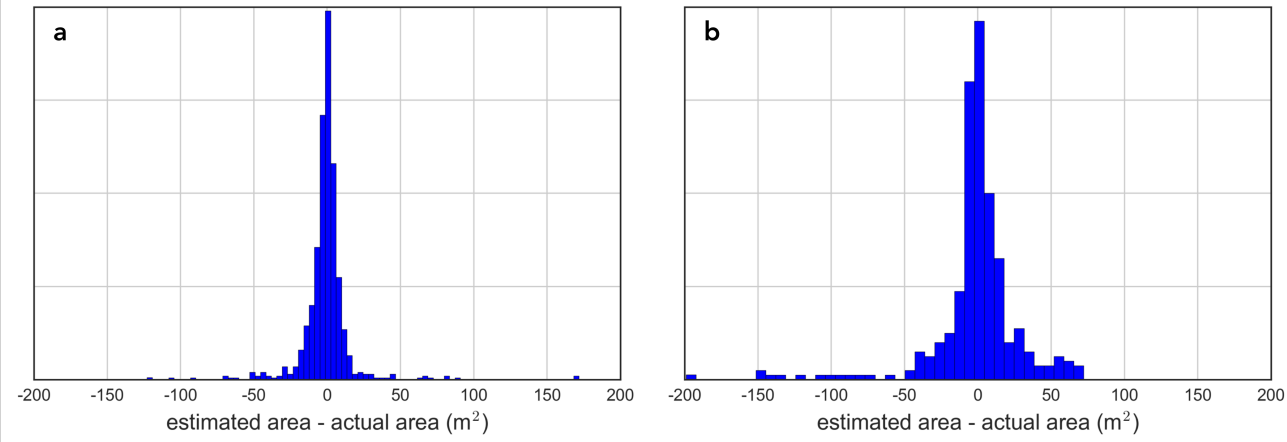

Figure S9: The distribution of area estimation residual at image tile level

The area estimation is nearly unbiased. **a.** For residential areas, the mean of residual is -0.949 m^2 and the standard deviation is 17.598 m^2. **b.** For non-residential areas, the mean of residual is -1.919 m^2 and the standard deviation is 31.331 m^2.

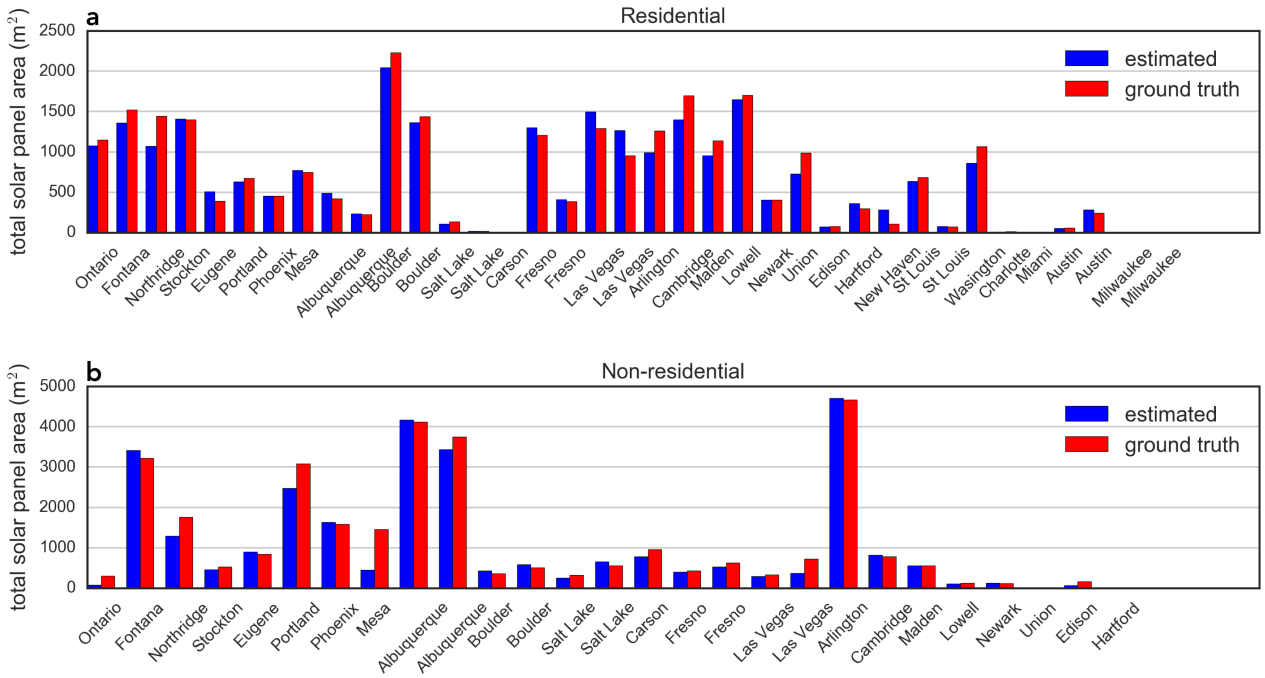

Figure S10: Region-level difference between estimated and actual total solar panel area

a. For residential areas. **b.** For non-residential areas. Some test regions have same name because they are in the same city but compiled from different districts (collection of census tracts).

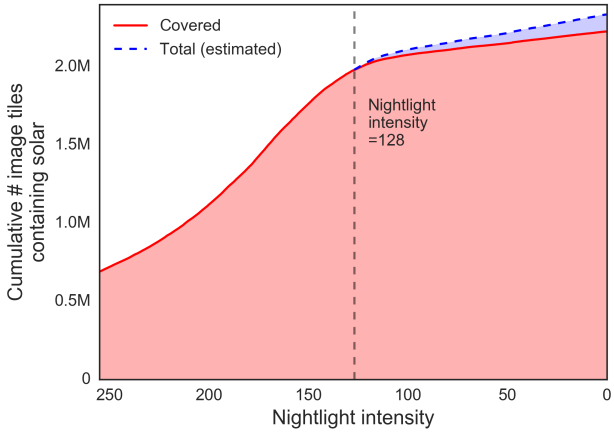

Figure S11: Cumulative number of image tiles containing solar panels as nightlight intensity *decreases* from 255 to 0 (covered versus total)

All urban areas as well as non-urban areas with nightlight intensity greater than 128 out of 255 are covered. Non-urban areas with nightlight intensity less than 128 will continue to be processed to complete the database. The number of image tiles with solar panels not covered in these regions are estimated utilizing the linear regression between the number of image tiles with solar panels in urban areas divided by the number of image tiles with solar panels, and nightlight intensity. The analysis determines that 95% of image tiles that contain a solar panel are covered. The difference between red line and blue dashed line represents the estimated number of solar panels not covered in the detected range.

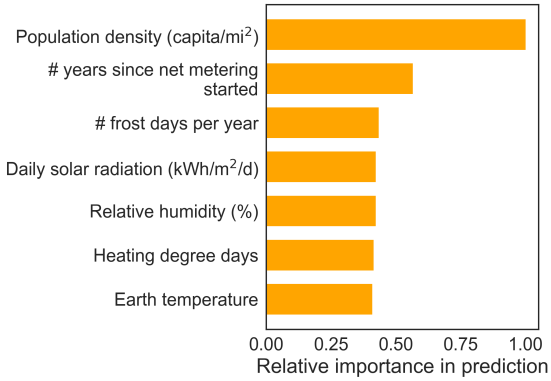

Figure S12: Relative feature importance of SolarNN

The importance of a feature input in SolarNN is calculated by averaging over the absolute values of gradients of output with respect to that input. All features are normalized before being fed into the neural network, thus the gradients of output with respect to each feature reflects the sensitivity of output to the variation of that feature. Highest feature importance is normalized to be 1.

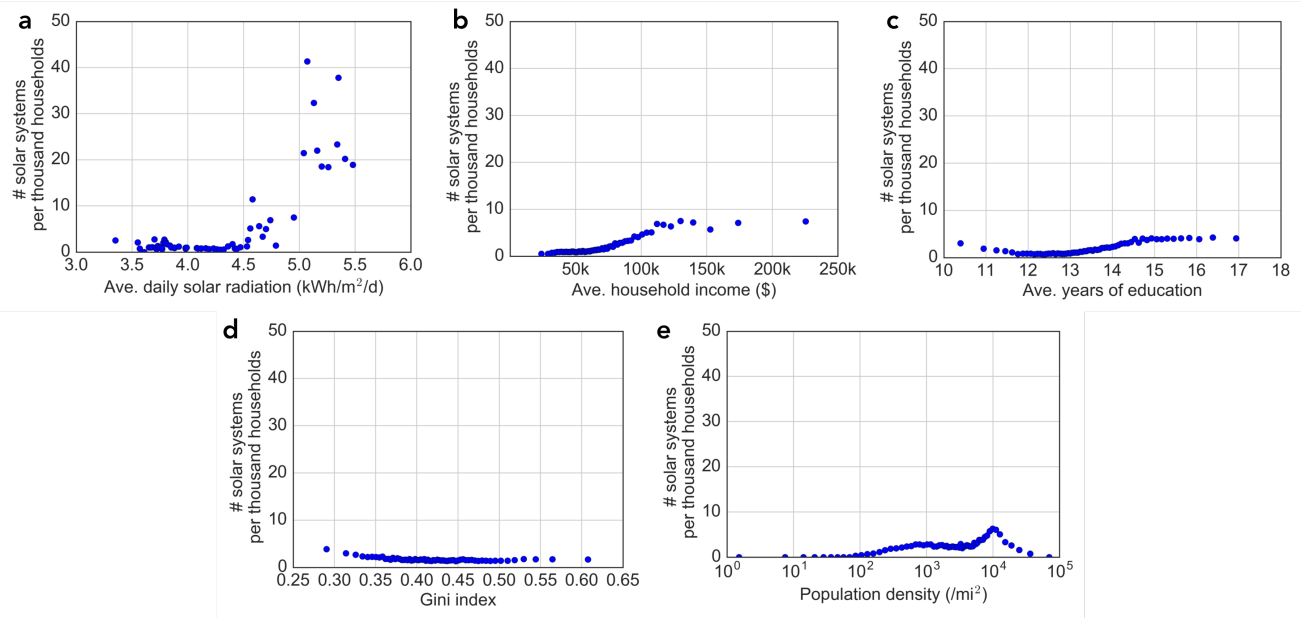

Figure S13: Basic trends between residential solar deployment density and different factors

a. For daily solar radiation; **b.** For average annual household income; **c.** For average years of education; **d.** For Gini index; **e.** For population density. All census tracts are grouped into 64 bins according to the value of the independent variable, for example, average daily solar radiation. For each bin, the median of the independent variable and the median of solar deployment density is plotted. All trends are plotted with the same scale for solar deployment density, showing that average daily solar radiation contributes to a significantly higher variability in solar deployment density compared to other factors.

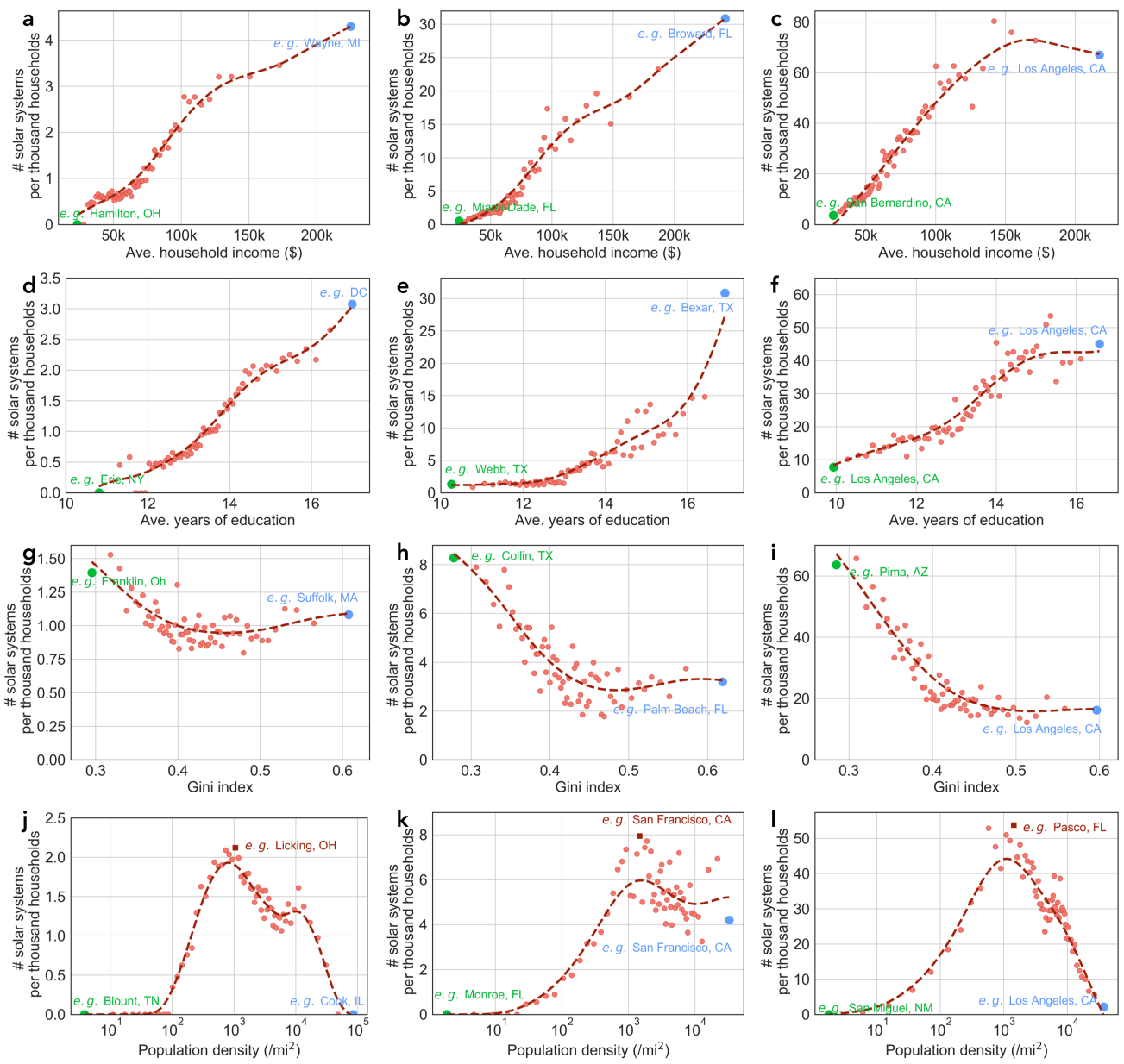

Figure S14: Correlation between solar deployment and average household income, average years of education, Gini index, population density, conditional on solar radiation

Fig. S13 defines the groupings. Blue/green/brown label denotes the county that the median census tract in the bin belongs to.

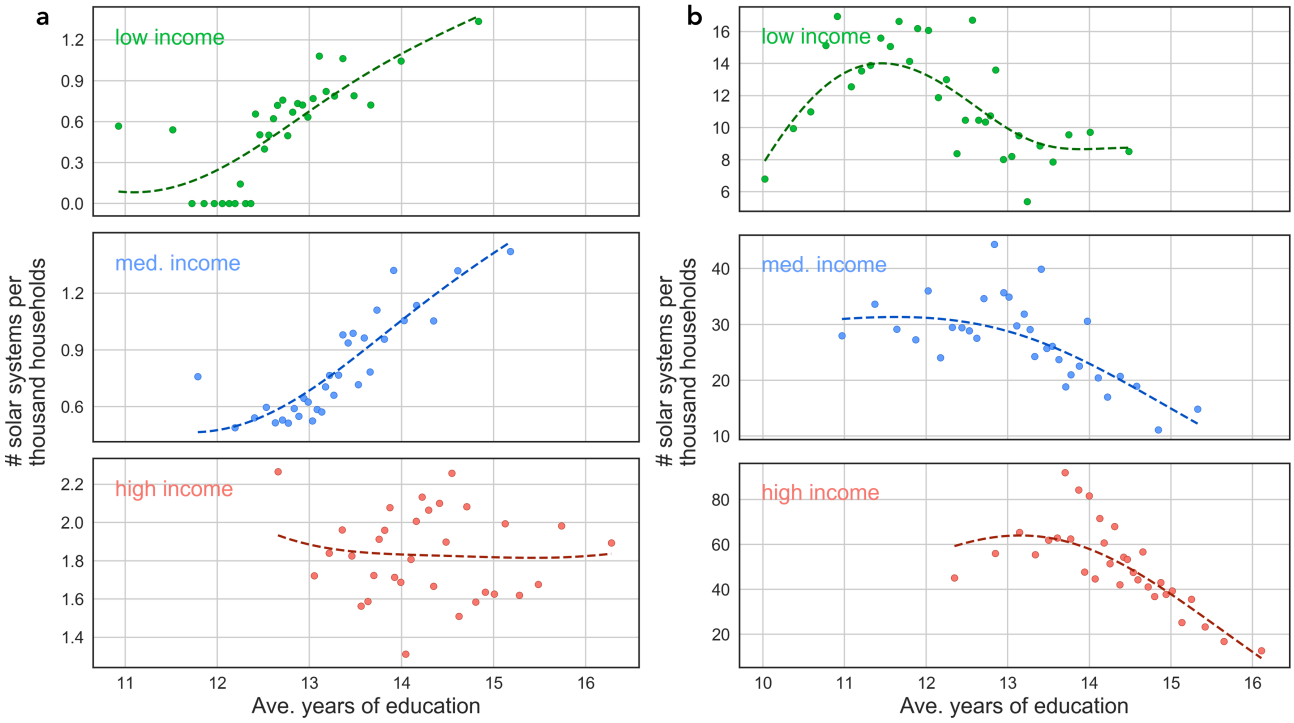

Figure S15: Education-solar deployment relationship conditional on average household income

For either high or low radiation level, data is divided into 7 groups according to income, and we select 2nd (low income), 4th (medium income), 6th (high income) group for visualization. **a.** In low radiation region (solar deployment not long-term profitable), solar deployment density increases with education level for low-income and medium-income group, but shows no correlation with education level for high-income group. **b.** In high radiation region (solar deployment is long-term profitable), solar deployment density does not increase (even decreases) with education level conditional on income. Fig. S13 defines the groupings.

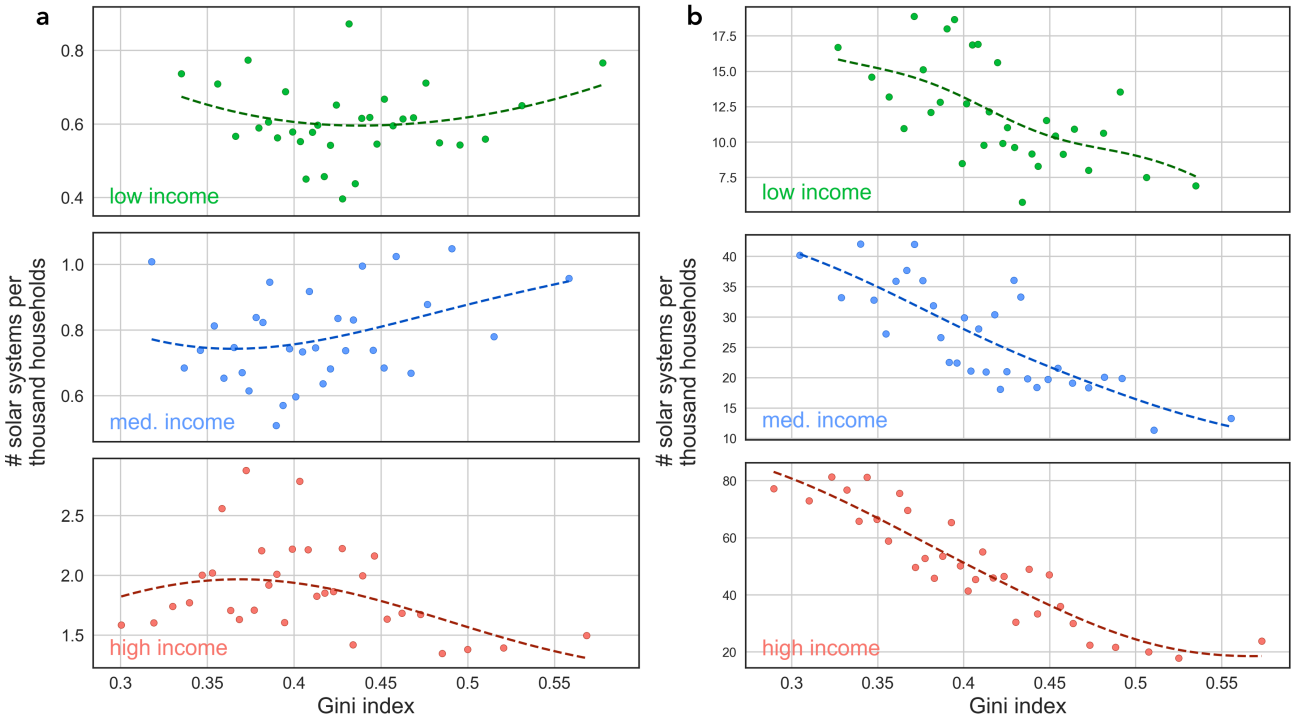

Figure S16: Gini index-Solar deployment relationship conditional on average household income

For each radiation/incentive/electricity price condition, data is divided into 7 groups according to income, and we select 2nd (low income), 4th (medium income), 6th (high income) group for visualization. a. In low radiation region (solar deployment not long-term profitable), solar deployment density shows no correlation with Gini index conditional on income. b. In high radiation region (solar deployment is long-term profitable), solar deployment density decreases with Gini index conditional on income. Fig. S13 defines the groupings.

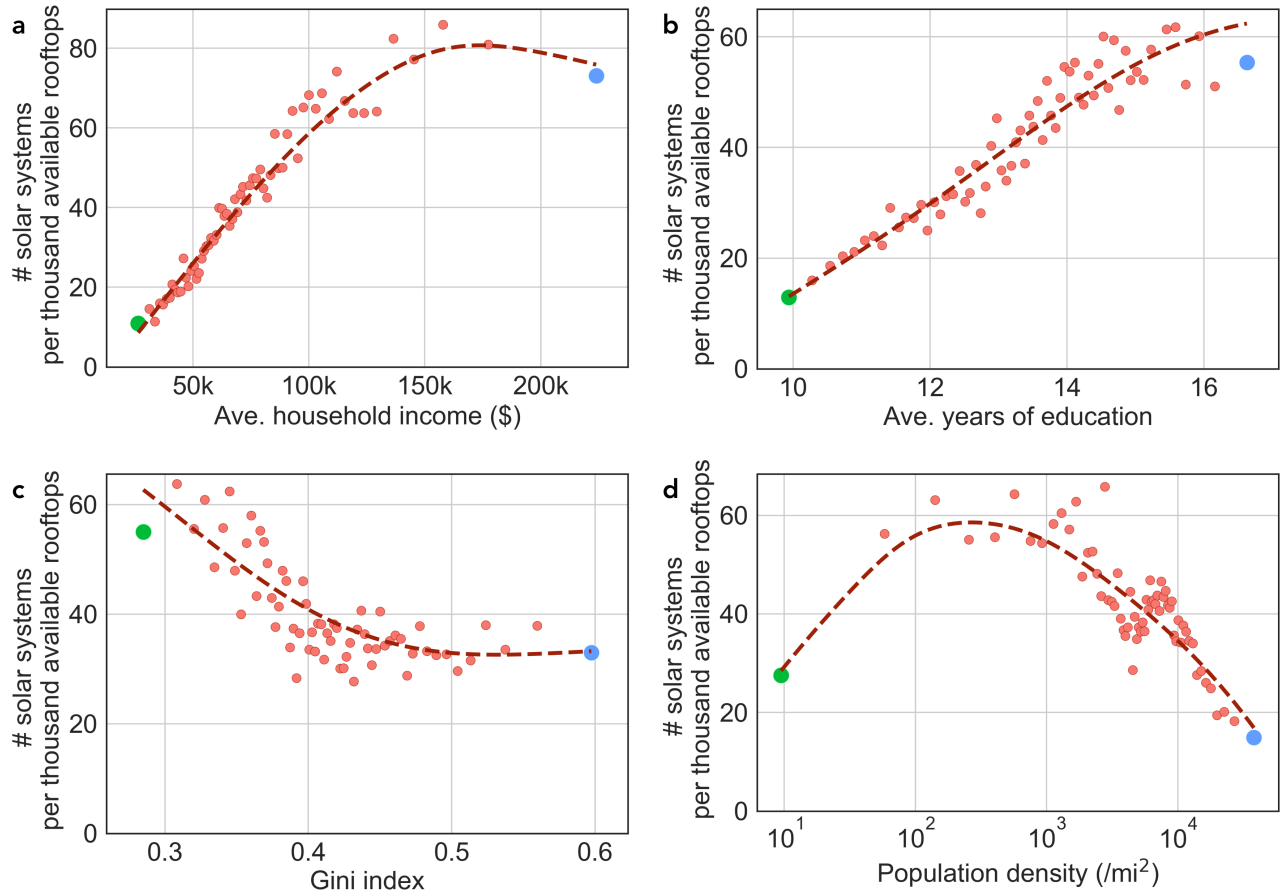

Figure S17: Solar deployment density measured by number of available rooftops correlates with demographic factors, conditional on solar radiation

We use census tracts with available rooftop data for analysis. The dashed lines are fitted with locally weighted scatterplot smoothing (LOWESS). Here we only show the correlations in high-radiation regions. **a.** Solar deployment density increases with average household income but saturates with high income. **b.** Solar deployment density increases with the average years of education. **c.** Solar deployment density decreases with the Gini index, but slope is flatter with higher Gini index. **d.** Solar deployment density decreases with population density when population density is high. Fig. S13 defines the groupings.

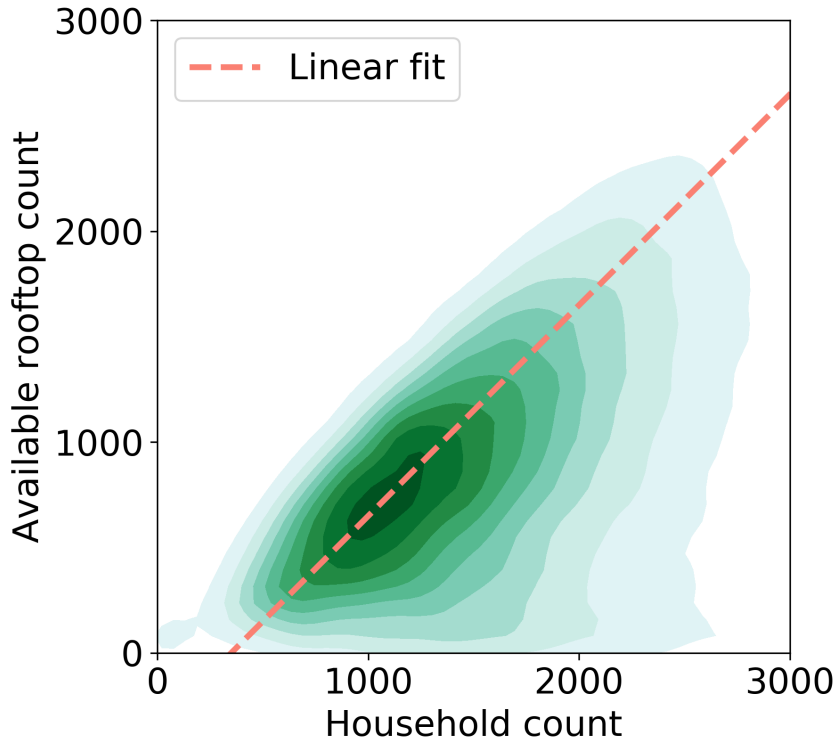

Figure S18: Correlation between household count and available rooftop count on census tract level

The x-axis is the household count, the y-axis is the available rooftop count for selected census tracts. The green contour plot is the density plot of the scatter of household count versus rooftop count for census tracts from Google's Sunroof Project. The red dashed line is a linear fit of the density plot that has equation $y = 0.999x - 333$.

Table S1. Dataset statistics

	Positive samples	Negative samples	Total samples
Training set	46,090	320,377	366,467
Validation set	226	12,760	12,986
Test set	1,221	92,279	93,500

Table S2. Typical years of education for different highest degree holders

Degree	Typical years of education
Middle school	8
High school	12
Junior college	14
Bachelor	16
Master	18
Professional degree	21
Doctoral degree	21

Table S3. Confusion matrix on test set (for residential areas).

	Actual positive	Actual negative
Predicted positive	774	57
Predicted negative	101	69,068

The threshold probability is 0.5, which is used in our work.

Table S4. Confusion matrix on test set (for non-residential areas).

	Actual positive	Actual negative
Predicted positive	313	21
Predicted negative	33	23,383

The threshold probability is 0.5, which is used in our work.

Algorithm 1 Searching adjacent solar panel pieces and merging

```
function FindAdjacent(tile, cluster, S)                    ▷ S - the set of positive tiles
    if not tile in S then return cluster.
    end if
    add tile to cluster
    remove tile from S
    if solar panel area in tile reach north edge then
        if northernTile in S and solar panel area in northernTile reach south edge then
            cluster = FindAdjacent(northernTile, cluster, S)    ▷ northernTile - the northern
adjacent tile of tile
        end if
    end if
    if solar panel area in tile reach south edge then
        if sorthernTile in S and solar panel area in southernTile reach north edge then
            cluster = FindAdjacent(southernTile, cluster, S)
        end if
    end if
    if solar panel area in tile reach east edge then
        if easternTile in S and solar panel area in easternTile reach west edge then
            cluster = FindAdjacent(easternTile, cluster, S)
        end if
    end if
    if solar panel area in tile reach west edge then
        if westernTile in S and solar panel area in westernTile reach east edge then
            cluster = FindAdjacent(westernTile, cluster, S)
        end if
    end if
     return cluster
end function
for t in S do
    build an empty cluster
    add FindAdjacent(t, cluster, S) into ClusterList
end for
```

Algorithm 2 The predicting process of SolarForest

```
function SolarForest(X)
    c = RandomForestClassifier(X)
    if c = 0 (containing no solar) then
        y = 0
    else
        y = RandomForestRegressor(X)
    end if
     return y
end function
```

REFERENCES

1. Haegel, N.M., Margolis, R., Buonassisi, T., Feldman, D., Froitzheim, A., Garabedian, R., Green, M., Glunz, S., Henning, H.M., Holder, B., and Kaizuka, I. (2017). Terawatt-scale photovoltaics: Trajectories and challenges. Science *356*,141-143.
2. Chu, S., and Majumdar, A. (2012). Opportunities and challenges for a sustainable energy future. Nature *488*, 294-303.
3. Agnew, S., and Dargusch, P. (2015). Effect of residential solar and storage on centralized electricity supply systems. Nature Climate Change *5*, 315-318.
4. National Renewable Energy Laboratory. The Open PV Project. https://openpv.nrel.gov.
5. Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. Science *353*, 790-794.
6. Malof, J. M., Bradbury, K., Collins, L. M., and Newell, R. G. (2016). Automatic detection of solar photovoltaic arrays in high resolution aerial imagery. Applied Energy *183*, 229–240.
7. Yuan, J., Yang, H.-H. L., Omitaomu, O. A., and Bhaduri, B. L. (2016). Large-scale solar panel mapping from aerial images using deep convolutional networks. Proceedings of the IEEE International Conference on Big Data, 2703–2708.
8. Malof, J. M., Bradbury, K., Collins, L. M., Newell, R. G., Serrano, A., Wu, H., and Keene, S. (2016). Image features for pixel-wise detection of solar photovoltaic arrays in aerial imagery using a random forest classifier. Proceedings of the IEEE International Conference on Renewable Energy Research and Applications, 799-803.
9. LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. Nature *521*, 436–444.
10. Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 248-255.
11. Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 1097–1105.
12. Pan, S. J. & Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering *22*, 1345–1359.
13. Malof, J. M., Collins, L. M., Bradbury, K., and Newell, R. G. (2016). A deep convolutional neural network and a random forest classifier for solar photovoltaic array detection in aerial imagery. Proceedings of the IEEE International Conference on Renewable Energy Research and Applications, 650–654.
14. Schaffer, A. J., and Brun, S. (2015). Beyond the sun—socioeconomic drivers of the adoption of small-scale photovoltaic installations in Germany. Energy Research & Social Science *10*, 220-227.
15. Kwan, C. L. (2012). Influence of local environmental, social, economic and political variables on the spatial distribution of residential solar PV arrays across the United States. Energy Policy *47*, 332–344.
16. Crago, C., and Chernyakhovskiy, I. (2014). Solar PV Technology Adoption in the United States: An Empirical Investigation of State Policy Effectiveness. Proceedings of the Agricultural & Applied Economics Association's Annual Meeting 27-29.
17. Rai, V. and McAndrews, K. (2012). Decision-making and behavior change in residential adopters of solar PV. Proceedings of the World Renewable Energy Forum.
18. Islam, T. and Meade, N. (2013). The impact of attribute preferences on adoption timing: The case of photovoltaic (PV) solar cells for household electricity generation. Energy Policy *55*, 521–530.
19. Vasseur, V., and Kemp, R. (2015). The adoption of PV in the Netherlands: A statistical analysis of adoption factors. Renewable and Sustainable Energy Reviews *41*, 483–494.
20. Palm, A. (2016). Local factors driving the diffusion of solar photovoltaics in Sweden: a case study of five municipalities in an early market. Energy Research & Social Science *14*, 1–12.
21. Rai, V., Reeves, D. C., and Margolis, R. Overcoming barriers and uncertainties in the adoption of residential solar PV. Renewable Energy *89*, 498–505.
22. Wolske, K. S., Stern, P. C., and Dietz, T. (2017). Explaining interest in adopting residential solar photovoltaic systems in the United States: toward an integration of behavioral theories. Energy Research & Social Science *25*, 134–151.
23. Braito, M., Flint, C., Muhar, A., Penker, M., and Vogel, S. (2017). Individual and collective socio-psychological patterns of photovoltaic investment under diverging policy regimes of Austria and Italy. Energy Policy *109*, 141–153.
24. Davidson, C., Drury, E., Lopez, A., Elmore, R., and Margolis, R. (2014). Modeling photovoltaic diffusion: An analysis of geospatial datasets. Environmental Research Letters *9*, 074009.
25. Letchford, J., Lakkaraju, K., and Vorobeychik, Y. (2014). Individual household modeling of photovoltaic adoption. AAAI Fall Symposium Series.
26. Li, H. and Yi, H. (2014). Multilevel governance and deployment of solar PV panels in us cities. Energy Policy *69*, 19–27.
27. De Groote, O., Pepermans, G., and Verboven, F. (2016). Heterogeneity in the adoption of photovoltaic systems in Flanders. Energy Economics *59*, 45–57.

28. Dharshing, S. (2017). Household dynamics of technology adoption: A spatial econometric analysis of residential solar photovoltaic (PV) systems in Germany. Energy Research & Social Science *23*, 113–124.
29. Breiman, L. (2001). Random forests. Machine Learning *45*, 5–32.
30. Bradbury, K., Saboo, R., Johnson, T.L., Malof, J.M., Devarajan, A., Zhang, W., Collins, L.M., and Newell, R.G. (2016). Distributed solar photovoltaic array location and extent dataset for remote sensing object identification. Scientific Data *3*, 160106.
31. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2818–2826.
32. Elkan, C. (2001). The foundations of cost-sensitive learning. International Joint Conference on Artificial Intelligence *17*, 973–978.
33. He, H. and Garcia, E. A. (2009). Learning from imbalanced data. Proceedings of the IEEE Transactions on Knowledge and Data Engineering *21*, 1263–1284.
34. Ling, C. and Sheng, V. (2009). Cost-sensitive learning and the class imbalance problem. Encyclopedia of Machine Learning: Springer.
35. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, *2921–2929*.
36. Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. Proceedings of the International Conference on Machine Learning, 807–814.
37. Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations.
38. Sarzynski, A., Larrieu, J., and Shrimali, G. (2012). The impact of state financial incentives on market deployment of solar technology. Energy Policy *46*, 550–557.
39. Official Energy Statistics from the U.S. Government. U.S. Energy Information Administration (EIA). https://www.eia.gov. Accessed: 2018-01-01.
40. 2011-2015 ACS 5-year estimates. https://www.census.gov/programs-surveys/acs/news/data-releases/2015/release.html. Accessed: 2017-07-01.
41. 2012 United States President Election Results. https://www.theguardian.com/news/datablog/2012/nov/07/us- 2012-election-county-results-download. Accessed: 2017-07-01.
42. 2016 United States President Election Results. https://townhall.com/election/2016/president. Accessed: 2017-07-01.
43. NASA Surface Meteorology and Solar Energy. https://eosweb.larc.nasa.gov/sse/. Accessed: 2017-09-01.
44. Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 39, 2481-2495.
45. Noh, H., Hong, S., and Han, B. (2015). Learning deconvolution network for semantic segmentation. Proceedings of the IEEE International Conference on Computer Vision, 1520–1528.
46. Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 3431–3440.
47. United States Census 2010. https://www.census.gov/2010census. Accessed: 2017-07-01.
48. Mellander, C., Lobo, J., Stolarick, K., and Matheson, Z. (2015). Night-time light data: A good proxy measure for economic activity? PLOS One *10*, e0139779.
49. New full-hemisphere views of earth at night. https://www.nasa.gov/image-feature/new-full-hemisphere-views-of-earth-at-night. Accessed: 2017-07-01.
50. Schelly, C. (2014). Residential solar electricity adoption: What motivates, and what matters? A case study of early adopters. Energy Research & Social Science *2*, 183–191.
51. Reinsberger, K., Brudermann, T., Hatzl, S., Fleiß, E., and Posch, A. (2015). Photovoltaic diffusion from the bottom-up: Analytical investigation of critical factors. Applied Energy *159*, 178–187.
52. Bollinger, B. and Gillingham, K. (2012). Peer effects in the diffusion of solar photovoltaic panels. Marketing Science *31*, 900–912.
53. Graziano, M. and Gillingham, K. (2014). Spatial patterns of solar photovoltaic system adoption: The influence of neighbors and the built environment. Journal of Economic Geography *15*, 815–839.