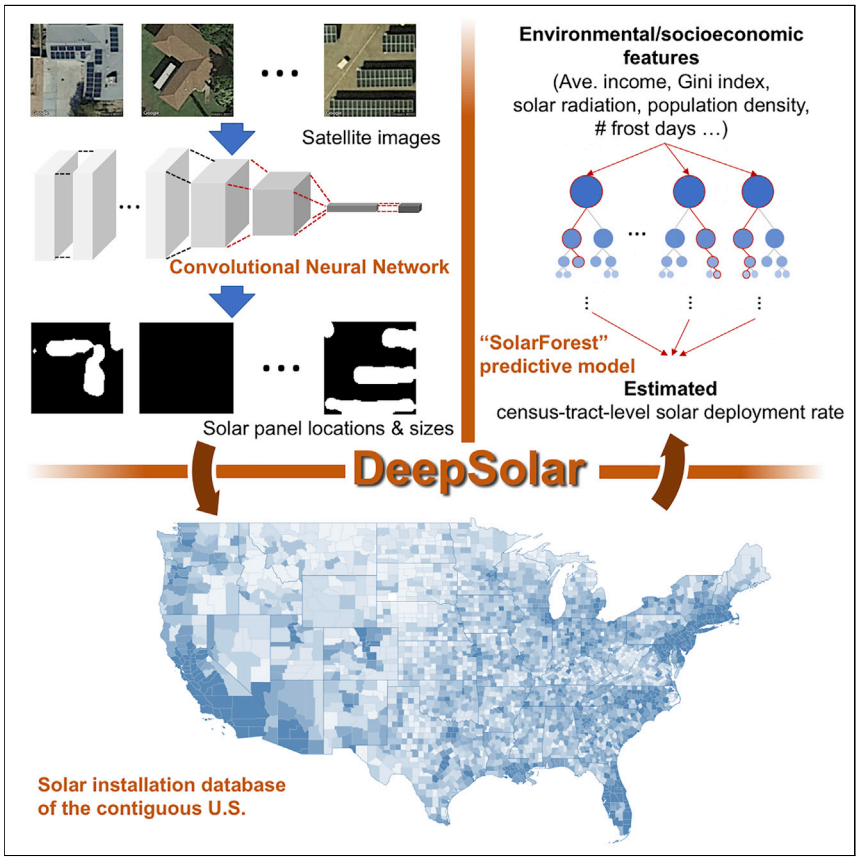
Joule

CellPress

Article

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



We developed an accurate deep learning framework to automatically localize solar photovoltaic panels from satellite imagery and estimate their sizes. We used it to construct a comprehensive and publicly available solar installation database of the contiguous US. We demonstrated its value by identifying key environmental and socioeconomic factors correlating with solar deployment, such as income and education. We also found that the solar deployment density can be accurately estimated at the microscopic level with these factors using a novel predictive model.

Jiafan Yu, Zhecheng Wang, Arun Majumdar, Ram Rajagopal

amajumdar@stanford.edu (A.M.)
ramr@stanford.edu (R.R.)

HIGHLIGHTS

An accurate deep learning model for detecting solar panel on satellite imagery

Built a nearly complete solar installation database for the contiguous US

Identified key socioeconomic factors correlating with solar deployment density

A predictive model to estimate solar deployment density at census tract level

Yu et al., Joule 2, 2605–2617
December 19, 2018
https://doi.org/10.1016/j.joule.2018.11.021

Joule

CellPress

Article

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

Jiafan Yu,[1,4] Zhecheng Wang,[2,3,4] Arun Majumdar,[2,*] and Ram Rajagopal[1,3,5,*]

SUMMARY

We developed DeepSolar, a deep learning framework analyzing satellite imagery to identify the GPS locations and sizes of solar photovoltaic panels. Leveraging its high accuracy and scalability, we constructed a comprehensive high-fidelity solar deployment database for the contiguous US. We demonstrated its value by discovering that residential solar deployment density peaks at a population density of 1,000 capita/mile^2, increases with annual household income asymptoting at ~$150k, and has an inverse correlation with the Gini index representing income inequality. We uncovered a solar radiation threshold (4.5 kWh/m^2/day) above which the solar deployment is "triggered." Furthermore, we built an accurate machine learning-based predictive model to estimate the solar deployment density at the census tract level. We offer the DeepSolar database as a publicly available resource for researchers, utilities, solar developers, and policymakers to further uncover solar deployment patterns, build comprehensive economic and behavioral models, and ultimately support the adoption and management of solar electricity.

INTRODUCTION

Deployment of solar photovoltaics (PVs) is accelerating worldwide due to rapidly reducing costs and significant environmental benefits compared with electricity generation based on fossil fuels.[1] Because of their decentralized and intermittent nature, cost-effective integration of solar panels on existing electricity grids is becoming increasingly challenging.[2,3] What is critically needed and currently unavailable is a comprehensive high-fidelity database of the precise locations and sizes of all solar installations. Recent attempts such as the Open PV Project[4] rely on voluntary surveys and self-reports. While they have been quite impactful in our understanding of solar deployment, they run the risk of being incomplete and with no guarantee on absence of duplication. Furthermore, with the rapid pace of solar deployment, such a database could become outdated. Machine learning combined with satellite imagery can be utilized to overcome the shortcoming of surveys.[5] The availability of satellite imagery with spatial resolution less than 30 cm for the majority of the US, which is annually updated, offers a rich data source for solar installation detection based on machine learning. Existing pixel-wise machine learning methods[6,7] suffer from poor computational efficiency, and relatively low precision and recall (cannot reach 85% simultaneously), while existing image-wise approaches[8] cannot provide system size or shape information. Google's Project Sunroof utilizes a proprietary machine learning approach to report locations without any size information. They have so far identified much

Context & Scale

We built a nearly complete solar installation database for the contiguous US utilizing a novel deep learning model applied to satellite imagery. The data are published as the first publicly available, high-fidelity solar installation database in the contiguous US. We plan to update it annually and add other countries and regions of the world. We demonstrated the value of this database by identifying key environmental and socioeconomic factors correlated with solar deployment. We also developed high-accuracy machine learning models to predict solar deployment density utilizing these factors as input. We hope the data produced by DeepSolar can aid researchers, policymakers, and the industry in gaining a better understanding of solar adoption and its impacts.

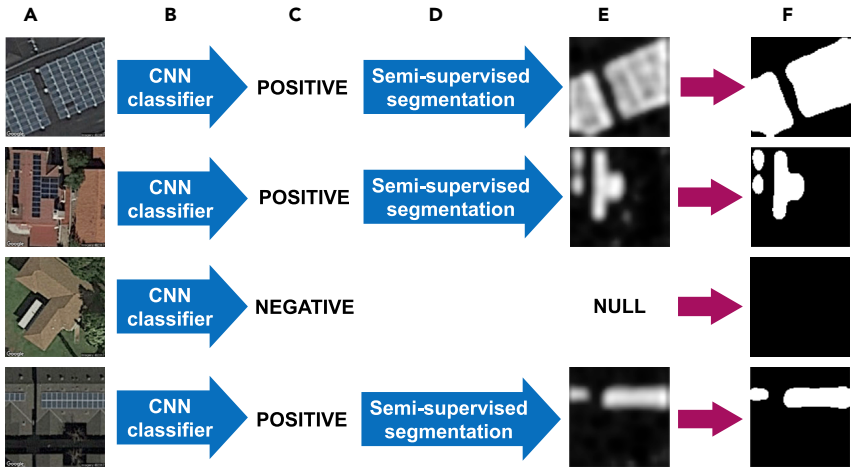


Figure 1. Schematic of DeepSolar Image Classification and Segmentation Framework

(A) Input satellite images are obtained from Google Static Maps.

(B) Convolutional neural network (CNN) classifier is applied.

(C) Classification results are used to identify images containing systems.

(D) Segmentation layers are executed on positive images and are trained with image-level labels rather than actual outlines of the solar panel, so it is "semi-supervised."

(E) Activation maps generated by segmentation layers where whiter pixels indicate higher likelihood of solar panel visual patterns.

(F) Segmentation is obtained applying a threshold to the activation map and finally both panel size and system counts can be obtained.

fewer systems (0.67 million) than in the Open PV database (~1 million) in the contiguous US.

Leveraging the development of convolutional neural networks (CNNs)[9] and large-scale labeled image datasets[10] for automatic image classification and semantic segmentation,[11] here we present an efficient and accurate deep learning framework called DeepSolar that uses satellite imagery to create a comprehensive high-fidelity database (which we called DeepSolar database) containing the GPS locations and sizes of solar installations in the contiguous US. To demonstrate the value of DeepSolar, we correlate environmental and socioeconomic factors with solar deployment data and have uncovered interesting trends with these factors. We utilize these insights to build SolarForest, the first high-accuracy machine learning predictive model that can estimate solar deployment density at the census tract level utilizing local environmental and socioeconomic features as input. We offer DeepSolar as a publicly available database that enables researchers to extract further insights about solar adoption, and aids policymakers to get deeper understanding and insights about socioeconomic and environmental correlations and causations.

RESULTS

Scalable Deep Learning Model for Solar Panel Identification

Generating a national solar installation database from satellite images requires a method that can learn to accurately identify panel location and size from very limited and expensive-to-obtain labeled imagery, while being computationally efficient to run at a nationwide scale. We developed DeepSolar, a novel semi-supervised deep learning framework featuring computational efficiency, high accuracy, and label-free training for size estimation (Figure 1). Traditionally, training a CNN to

[1]Department of Electrical Engineering, Stanford University, 450 Serra Mall, Stanford, CA 94305, USA

[2]Department of Mechanical Engineering, Stanford University, 450 Serra Mall, Stanford, CA 94305, USA

[3]Department of Civil & Environmental Engineering, Stanford University, 450 Serra Mall, Stanford, CA 94305, USA

[4]These authors contributed equally

[5]Lead Contact

*Correspondence: amajumdar@stanford.edu (A.M.), ramr@stanford.edu (R.R.)

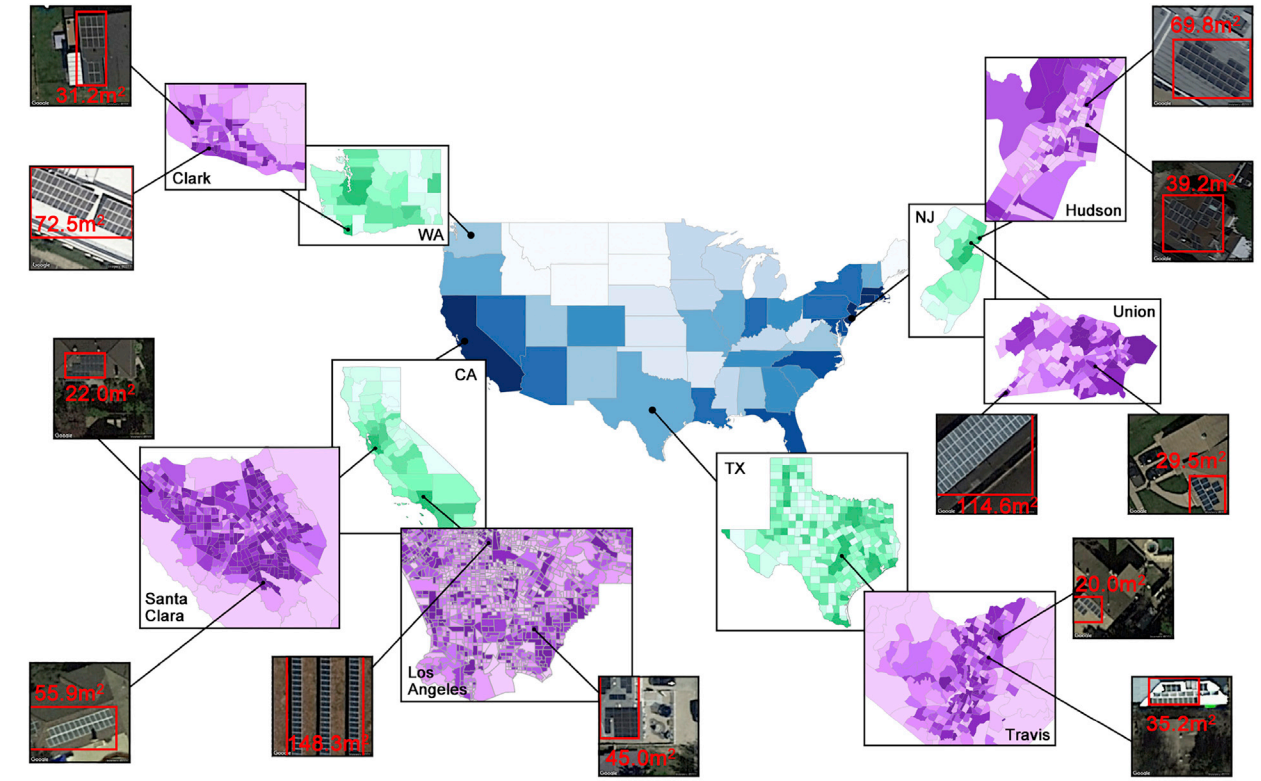
https://doi.org/10.1016/j.joule.2018.11.021

classify images requires massive training samples with true image-level class labels, while training it to segment objects requires large training set with ground truth pixel-wise segmentation annotations, which are extremely expensive to construct. Furthermore, fully supervised segmentation has relatively poor computation efficiency.[6,7] To enable efficient solar panel identification and segmentation, DeepSolar first utilizes transfer learning[12] to train a CNN classifier on 366,467 images sampled from over 50 cities/towns across the US with merely image-level labels indicating the presence or absence of panels. Segmentation capability is then enabled by adding an additional CNN branch directly connected to the intermediate layers of the classifier, which is trained on the same dataset to greedily extract visual features to generate clear boundaries of solar panels without any supervision of actual panel outlines. Such a "greedy layer-wise training" technique greatly enhances the semi-supervised segmentation capability, making its performance comparable with fully supervised methods. The output of this network is an activation map that involves a threshold to produce panel outlines. Segmentation is not applied on samples predicted to contain no panels, greatly enhancing the computation efficiency. Details can be found in Experimental Procedures and Supplemental Information.

The performance of our model is evaluated on a test set containing 93,500 randomly sampled images across the US. We utilize precision (rate of correct decisions among all positive decisions) and recall (ratio of correct decisions among all positive samples) to measure classification performance. DeepSolar achieves a precision of 93.1% with a recall of 88.5% in residential areas and a precision of 93.7% with a recall of 90.5% in non-residential areas. Such a result is significantly higher than previous reports.[6–8,13] Furthermore, our performance evaluation guarantees far more robustness since their test sets were only obtained from one or two cities but ours are sampled from nationwide imagery. Mean relative error (MRE), the area-weighted relative error, is used to measure size estimation performance. The MRE is 3.0% for residential areas and 2.1% for non-residential areas for DeepSolar. The errors are independent and nearly unbiased so MRE decreases even further when measured over larger regions. See more details in Supplemental Information Section 2.3.

Nationwide Solar Installation Database

DeepSolar was used to scan, within a month, over one billion image tiles covering all urban areas as well as locations with reasonable nighttime lights to construct the first complete solar installation profile of the contiguous US with exact locations and sizes of solar panels (see Supplemental Information Section 2.4 for details). The number of detected solar systems in the contiguous US is (1.4702 $\pm$ 0.0007) million, which exceeds the 1.02 million installations without accurate location in Open PV[4] and the 0.67 million installations without size information in Project Sunroof. In our detected installation profile, a solar system is a set of solar panels on top of a building, or at a single location such as a solar farm. We built a complete resource density map in the contiguous US from state level to household level (Figure 2). Solar installation densities have dramatic variability at state (e.g., 1.34–224.1 $m^2/mile^2$) and county levels (e.g., 255–7,490 $m^2/mile^2$ in California). Distributed residential-scale solar systems are 87% of the total system counts, but 34% of the total panel area in our database, and 23.4% of the census tracts contain 90% of the residential-scale installations (Figure 3A). Only 2,998 census tracts (4%) have more than 100 residential-scale systems (Figure 3B). The median of average system size for tracts with different levels of residential



state level	county level				census tract level					
contiguous U.S.	CA	TX	NJ	WA	Santa Clara	Los Angeles	Travis	Union	Hudson	Clark
<1.34	<0.69	<0.03	<132	<0.11	< 255	< 618	< 70	< 323	< 11	< 0.84
1.34 - 3.45	0.69 - 6.53	0.03 - 0.15	132 -507	0.11 - 0.27	255 - 921	618 - 1181	70 - 205	323 - 598	11 - 327	0.84 - 22.53
3.45 - 4.34	6.53 - 21.18	0.15 - 0.27	507 - 729	0.27 - 0.51	921 - 1526	1181 - 1579	205 - 301	598 - 1092	327 - 743	22.53 - 49.59
4.34 - 7.92	21.18 - 90.44	0.27 - 0.46	729 - 1194	0.51 - 1.33	1526 - 2152	1579 - 2027	301 - 435	1092 - 1499	743 - 1460	49.59 - 79.60
7.92 - 14.24	90.44 - 165.9	0.46 - 1.03	1194 - 1325	1.33 - 1.90	2152 - 2864	2027 - 2531	435 - 647	1499 - 2022	1460 - 3331	79.60 - 95.80
14.24 - 24.12	165.9 - 281.9	1.03 - 1.71	1325 - 1814	1.90 - 5.95	2864 - 3753	2531 - 3176	647 - 842	2022 - 2942	3331 - 6598	95.80 - 131.59
24.12 - 76.74	281.9 - 446.2	1.71 - 3.98	1814 - 3027	5.95 - 11.68	3753 - 5071	3176 - 4079	842 - 1319	2942 - 4398	6598 - 9971	131.59 - 165.94
76.74 - 224.1	446.2 - 1088.1	3.98 - 11.95	3027 - 4148	11.68 - 24.20	5071 - 7490	4079 - 6366	1319 - 2197	4398 - 7960	9971 - 38058	165.94 - 281.21
> 224.1	> 1088.1	> 11.95	> 4148	> 24.20	> 7490	> 6366	> 2197	> 7960	> 38058	> 281.21

Figure 2. Solar Resource Density (Solar Panel Area per Unit Area [m^2/$mile^2$]) at State, County, and Census Tract Levels, with Examples of Detected Solar Panels

Darker colors represent higher solar resource density. Several census tracts in Hudson County, New Jersey, have solar resource density higher than 30,000 m^2/$mile^2$ while the five northern states (Montana, Idaho, Wyoming, North Dakota, and South Dakota) have solar resource density less than 1.34 m^2/$mile^2$, indicating extremely heterogeneous spatial distributions. The red-line rectangles denote the predicted bounding boxes of solar power systems in image tiles and the values denote the estimated area of solar systems.

solar system counts are all between 20 and 27 m^2 (Figure 3B). Due to the distributed nature of residential solar systems and their small variability in sizes, in this work we focus on residential solar deployment density defined as the number of residential-scale systems per thousand households at the census tract level. Leveraging our database, non-residential solar deployment can also be extensively analyzed in the future.

Correlation between Solar Deployment and Environmental/Socioeconomic Factors

We correlate the residential solar deployment with environmental factors such as solar radiation and socioeconomic factors from US census data to uncover solar deployment trends. We also collect and consider possible financial indicators reflecting the cumulative effects of energy policies, including the average electricity

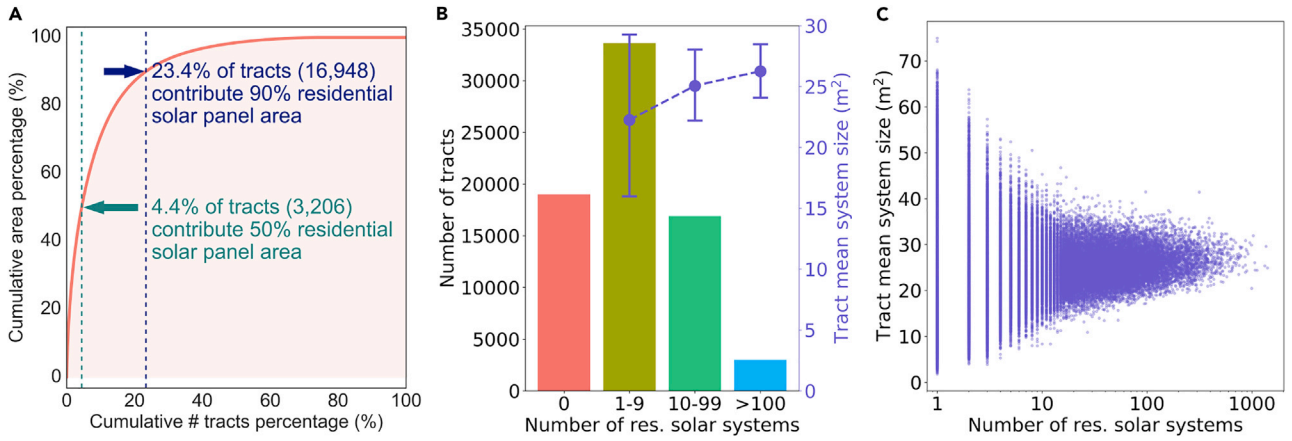


Figure 3. Residential Solar Deployment Statistics at Census Tract Level

(A) Cumulative distribution of residential solar area over census tracts.

(B) Tract counts of the number of solar systems. The left y axis is the number of census tracts, corresponding to the bar plot. The right y axis is the tract mean solar system size, corresponding to the purple error bar plot. The purple dots are the medians of tract average system size within each category; the error bars represent 25% and 75% percentiles.

(C) Mean system size of a tract varies with the number of residential solar systems in the tract. Each point represents one census tract. When the number of system increases, the mean size converges to 25 m^2.

retail rate over the past 5 years, the number of years since the start of net metering, and other types of financial incentives.

Results show that solar deployment density sharply increases when solar radiation is above 4.5–5 kWh/m^2/day (Figure 4A), which we define as an "activation" threshold triggering the increase of solar deployment. When we dissect this trend according to electricity rates (Figure 4C), we find that the activation threshold is clear for low-electricity-rate regions, but it is unclear in high-electricity-rate regions, indicating that this threshold may reflect a potential financial break-even point for deep penetration of solar deployment.

Since significant variation of solar deployment density is observed with solar radiation (see Supplemental Information Section 3.1 for details), we split all tracts into three groups according to the radiation levels (low, medium, and high), and analyze the trends with other factors based on such grouping. Population/housing density has been observed to be positively[14] or negatively[15,16] correlated with solar deployment. Figure 5A shows that both trends hold but with a peak deployment density at the population density of 1,000 capita/$mile^2$. Rooftop availability is not the limiting factor as the trend persists when we compute the number of systems per thousand rooftops (see Supplemental Information Section 3.2). Annual household income is a substantial driver for solar deployment (Figure 5B). Low- and medium-income households have low deployment densities despite solar systems being profitable for high-radiation rates, indicating that the lack of financial capability of covering the upfront cost is likely a major burden of solar deployment. Surprisingly, we observe that the solar deployment in high-radiation regions saturates at annual household incomes higher than $150,000 indicating other limiting factors. Solar deployment density rate also shows an increasing trend with average education level (Figure 5C). However, if conditioning on income, this trend actually does not hold in regions with high radiation, but still holds in the regions with poor solar radiation and lower income level (Figure S15). Moreover, solar deployment density in census tracts with high radiation is strongly correlated, and decreasing, with

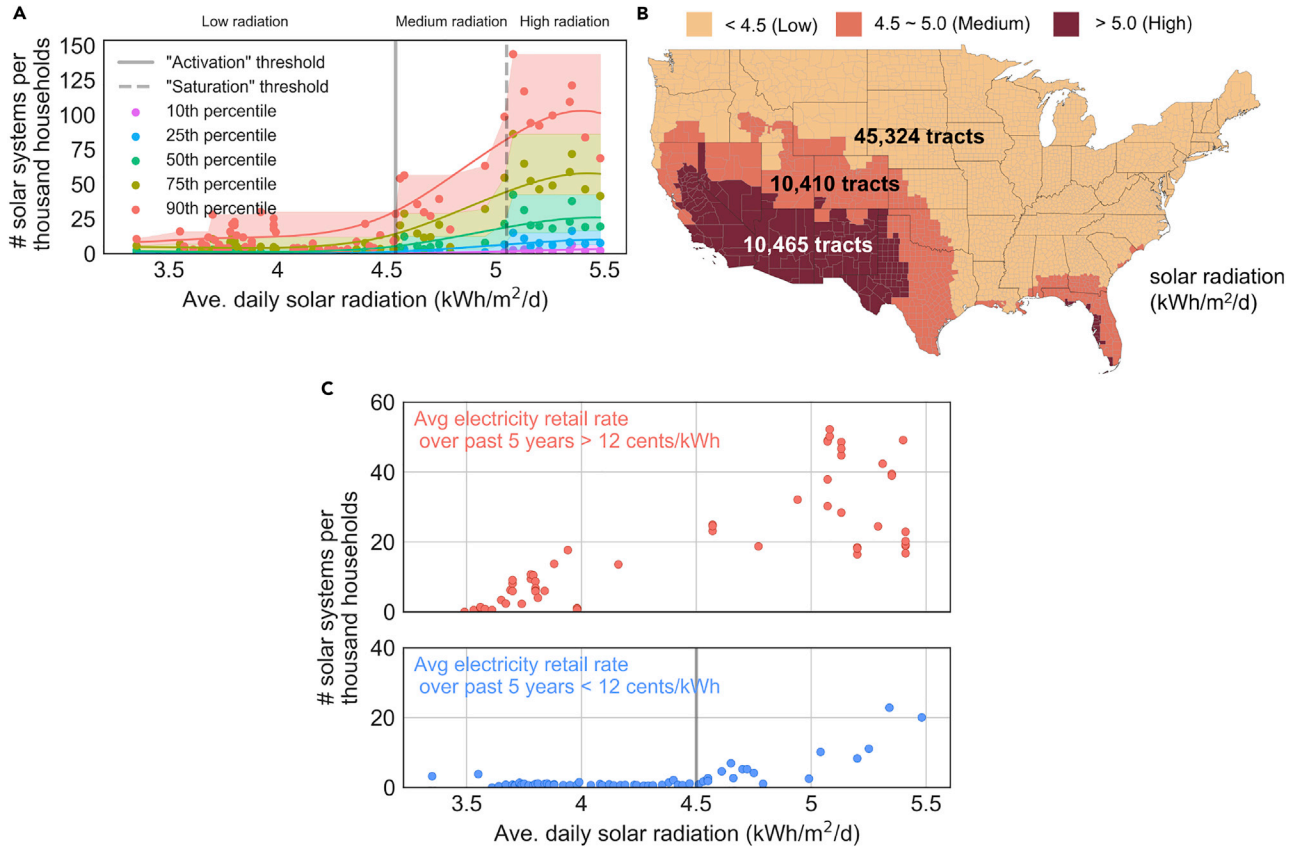


Figure 4. Correlation between Solar Radiation and Solar Deployment

(A) Solar deployment density has non-linear relationship with solar radiation. Two thresholds (4.5 and 5.0 kWh/m^2/day) are observed for all percentiles. Shaded areas represent the cumulative maximum of percentile scatters. Census tracts are grouped according to 64 bins of solar radiation. Curves are fitted utilizing locally weighted scatterplot smoothing (LOWESS).

(B) US map colored according to the three levels of average solar radiation defined by the thresholds identified in (A).

(C) Solar deployment density correlation with solar radiation, conditioning on the level of electricity retail rate.

the Gini index, a measure of income inequality (Figure 5D). Additional trends that illustrate racial and cultural disparities, for example, can be extracted utilizing this database. We expect that routinely updating the DeepSolar large-scale database and making it publicly available can empower the community to uncover further insights.

Predictive Solar Deployment Model

Models that estimate deployments from socioeconomic and environmental variables are key for decision making by regulatory agencies, solar installers, and utilities. Studies have focused on either utilizing surveys[17–23] or data-driven approaches[14–16,24–28] at spatial scales ranging from county- to state-level models, achieving in-sample R^2 values between 0.04 and 0.71. The models are typically linear[28] or log-linear[27] and utilize less than 10,000 samples for regression. Our result instead reveals that socioeconomic trends are highly non-linear. Furthermore, our database, generated by DeepSolar, offers abundant data points to develop elaborate non-linear models. Hence, we build and compare several accurate predictive models to estimate solar deployment at census tract level utilizing the data from more than 70,000 census tracts (see details in Experimental Procedures). Each model takes 94 environmental and socioeconomic

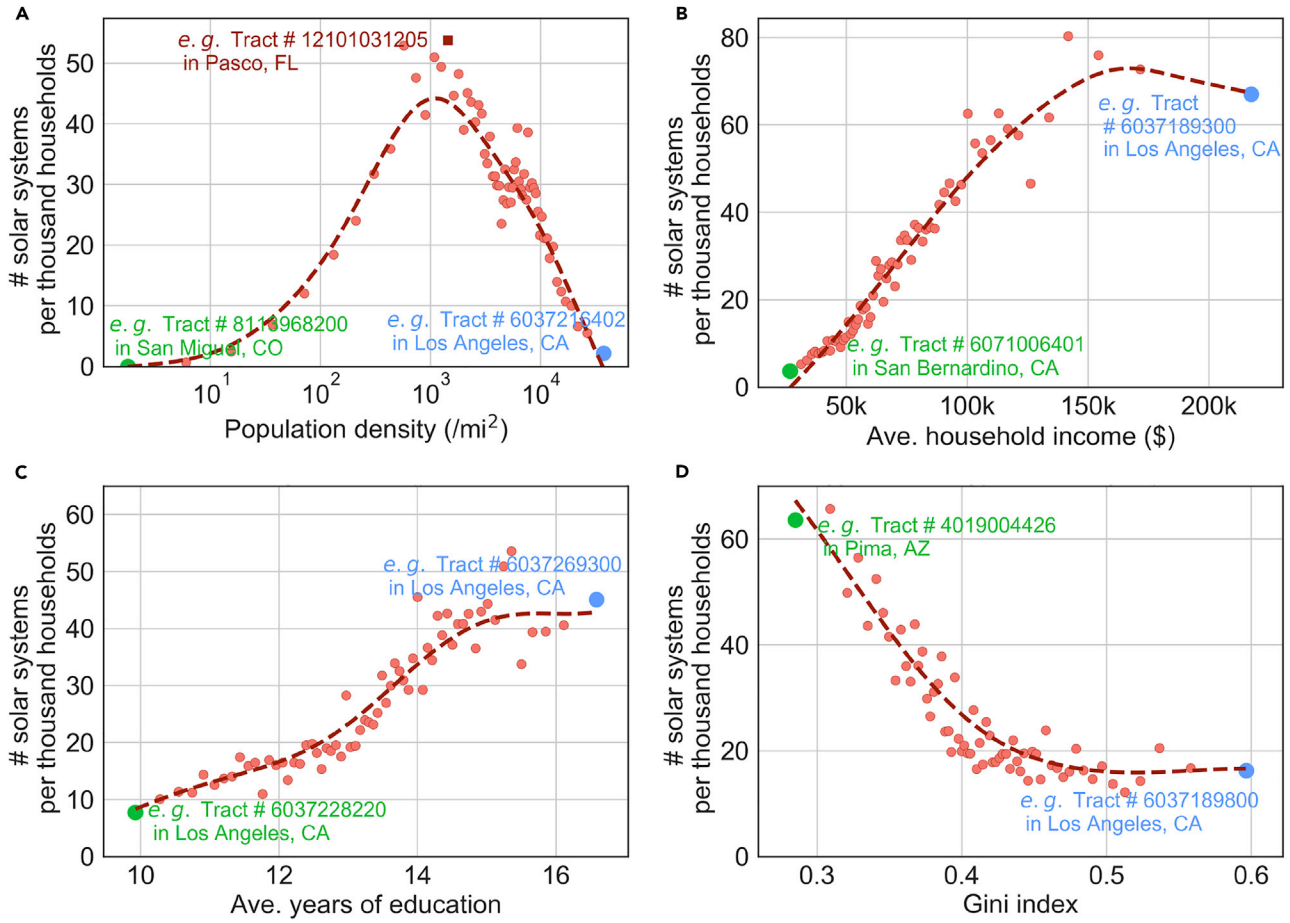


Figure 5. Residential Solar Deployment Density Correlates with Socioeconomic Factors Conditional on Radiation

Census tracts are grouped according to 64 bins of the target factor. Curves are fitted utilizing LOWESS. Blue/green/brown labels denote the county that the median census tract in the bin belongs to. Here we only show tracts with high solar radiation (>5.0 kWh/m^2/day). Complete trends are shown in Figure S14.

(A) Solar deployment density increases with population density with a peak at 1,000 capita/mile^2.

(B) Solar deployment density increases with average annual household income but saturates at incomes of $150k.

(C) Solar deployment density increases with the average years of education.

(D) Solar deployment density decreases with income inequality in a tract and a critical Gini index of 0.4 saturates solar deployment.

factors as inputs, such as solar radiation, average electricity retail rate over past 5 years, number of years since the start of different types of financial incentives, average household income, etc. (see details in Supplemental Information Section 1.2). These 94 factors are the largest set of factors we can collect for all census tracts and part of them have been also utilized and reported in previous works.[14–16,24–28]

Among all predictive models, the Random Forest-based model, called SolarForest, achieves the tier-1 out-of-sample R^2 value of 0.72 in the 10-fold cross-validation, which is even higher than the in-sample R^2 values of any other models in previous works.[14–16,24–26,28] SolarForest is a novel machine learning-based hierarchical predictive model that postulates census tract level solar deployment as a two-stage process: whether tracts contain solar panels or not, and, if they do contain them, the number of systems per household is decided (Figure 6A). Each stage utilizes a Random Forest[29] that takes all 94 factors. By ranking the

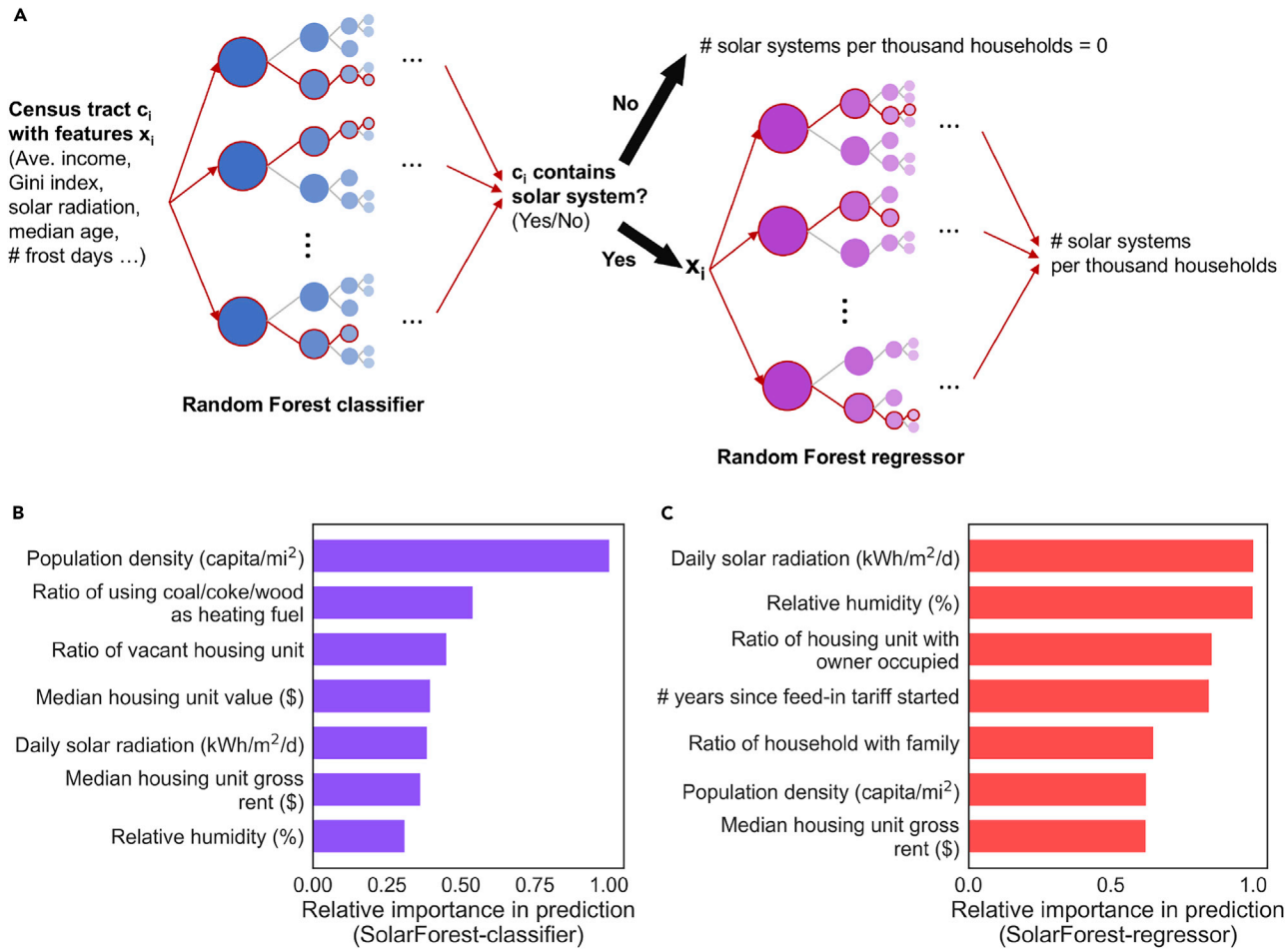


Figure 6. Architecture and Feature Importance of SolarForest

(A) SolarForest combines two random forests: a Random Forest binary classifier (blue) to predict whether a census tract contains at least one solar system, and a Random Forest regression model (magenta) to estimate the number of solar systems per thousand households if the tract contains solar systems. The classifier consists of 100 decision trees and the regressor consists of 200 decision trees. Each circle in decision trees represents a node for binary partitioning according to the value of one feature. If the output of the classifier is "Yes," the final prediction of solar deployment density rate is the output of the regressor, else the solar deployment density is predicted to be zero.

(B) The relative feature importance in SolarForest model, classification stage.

(C) The relative feature importance in SolarForest model, regression stage.

feature importance in prediction at both stages, we observe that population density is the most significant feature to decide whether a census tract contains solar systems (Figure 6B); for a census tract containing solar systems, environmental features such as solar radiation, relative humidity, and number of frost days serve as the most important predictors to estimate the solar deployment density (Figure 6C).

DISCUSSION

DeepSolar is a novel approach to create, publish, update, and maintain a comprehensive open database on the location and size of solar PV installations. We aim to continuously update the database to generate a time-history of solar installations and increase coverage to include all of North America, including remote areas with utility-scale solar, and non-contiguous US states. Eventually the database will

include all regions in the world that have high-resolution imagery. In this work, we only estimated the horizontal projection areas of solar panels from satellite imagery. In the future, based on the existing GPS location information, we aim to continue using deep learning methods to infer roof orientation and tilt information from street view images, enabling more accurate estimation of solar system size and solar power generation capacity. In addition, the database is linked to US demographic data, solar radiation, utility rates, and policy information. We demonstrated that this rich database led to the discovery of previously unobserved non-linear socioeconomic trends of solar deployment density. It also enabled the development of state-of-the-art predictive models on solar deployment based on machine learning. As we update the database annually, such predictive models can be further improved to forecast the annual increment of solar installations in the census tracts according to the local environmental and socioeconomic factors. In the near future, this database can be utilized to develop granular adoption models relying on richer information on electricity rates and incentives, conduct causal inferences, and gain nuanced understanding of peer effects, inequality and other sociocultural trends in solar deployment. It can serve as a starting point to develop engineering models for solar generation in power distribution systems. The DeepSolar database closes a significant gap for the research and policy community, while at the same time advances methods in semi-supervised deep learning on satellite data and solar deployment modeling.

EXPERIMENTAL PROCEDURES

Massive Satellite Imagery Dataset

A massive amount of image samples is essential for developing a CNN model, since CNN can only gain good generalization ability with a large number of labeled samples for training. Bradbury et al.[30] built a manually labeled dataset based on US Geological Survey orthoimagery. However, it is sampled from only four cities in California, failing to cover the nationwide diversity, and thus they cannot guarantee the model developed with it to still perform well on other regions. In comparison, we have built a large-scale satellite image dataset based on the Google Static Map API with images collected to cover comprehensively the contiguous US (50 cities/towns). Our dataset consists of a training set (366,467 samples), a validation set (12,986 samples), and a test set (93,500 samples). The percentage of images in the dataset for model development compared with the total number of images we scanned so far in the US is 0.043%. Images in the test set are randomly sampled by generating random latitude and longitude within rectangular regions totally different from those in the training set. To train both classification and segmentation capabilities, an image-level label, indicating positive (containing solar panel) or negative (not containing solar panel), is annotated for all samples in the dataset. To evaluate the ability of size estimation, each test sample is also annotated with ground truth regions of solar panels beside image-level labels. We are also making this dataset public for the research community to drive model developing and testing on specific computer vision tasks. See Supplemental Information Section 1.1 for more details.

System Detection Using Image Classification

We utilize a state-of-art CNN architecture called Inception-v3[31] as our basic classification framework. The Inception-v3 model is pre-trained with 1.28 million images containing 1,000 different classes in the 2014 ImageNet Large Scale Visual Recognition Challenge,[10] and achieves 93.3% top 5 accuracy on that dataset. We start from the pre-trained model since the diversity from the massive dataset helps the CNN learn basic patterns of images across multiple domains. The model is then

developed on our training set by re-training the final affine layer from randomized initialized parameters and fine-tuning all other layers starting from the well-trained parameters. This process, called transfer learning,[12] is becoming common practice in deep learning and computer vision. The output of our model is a set of two probabilities indicating positive (containing solar) and negative (not containing solar).

The outputs of our model are two probabilities indicating positive (containing solar) and negative (not containing solar). The distribution of binary solar panel labels is extremely skewed in the training set (46,090 positive in 366,467 total) since solar panels are very rare compared with the whole territory. We solve this problem with a cost-sensitive learning framework,[32–34] which automatically sets more penalty to the misclassifications of positive samples than negative samples (see details in Supplemental Information Section 2.1).

Size Estimation Using Semi-supervised Segmentation

In addition to identifying whether an image tile contains solar panels, we also develop a semi-supervised method to accurately localize solar panels in images and estimate their sizes. Compared with fully supervised approaches suffering from low computation efficiency and requiring a large number of training samples with ground truth segmentation annotations, our semi-supervised segmentation model requires only image-level labeled (containing solar or not) images for training, which is achieved by greedily extracting visual patterns from intermediate results of classification. Roughly speaking, in CNN, the output of each convolutional layer is a stack of feature maps, each representing different feature activations. With the linear combination of these visual patterns, we can obtain a class activation map (CAM)[35] indicating the most activated regions of our target object, a solar panel. Furthermore, in CNN, features learned at upstream layers represent more general patterns such as edges and basic shapes, while features learned at downstream layers represent more specific patterns. As a result, upstream feature maps are more complete but noisy, while downstream feature maps are more discriminative but incomplete. By greedily extracting features at upstream layers, we can generate both complete and discriminative CAM for segmentation. To achieve that, we repeat greedily training a series of layers for classification and adding new layers after training (see details in Supplemental Information Section 2.2). Such a greedy layer-wise training mechanism is for the first time proposed for semi-supervised object segmentation. The code for system detection and size estimation is available here: http://web.stanford.edu/group/deepsolar/home.

Distinguish between Residential and Non-residential Solar

Our database contains both residential and non-residential solar panel data. We distinguish between residential and non-residential solar panels since they have different usages, scales, and economic natures. Due to the size, shape, and location differences of these two types of solar panels, we utilize a logistic regression model and train it with four basic features of each solar system: solar system area, nightlight intensity, the ratio between the solar system area and its bounding box area, and a Boolean, indicating if the system is merged from a single image tile. Since the non-residential solar systems only account for a small proportion, we also assign different weights, which are inversely proportional to the quantity ratio, to the misclassification of these two types during training. The training set size is 5,000 and the test set size is 1,078. Out-of-sample tests show that the recall is 81.3% for the residential type and 98.5% for the non-residential type, and the precision is 96.8% for residential type and 90.6% for non-residential types on the test set. These results are in terms of area.

Table 1. Comparison of the Cross-Validation R^2 Value of Different Solar Deployment Predictive Models

Model	Cross-Validation R^2
LR (quadratic + interaction)	0.181
MARS	0.267
RF regressor	0.412
RF classifier + LR (quadratic + interaction)	0.643
RF classifier + MARS	0.592
SolarForest (RF classifier + RF regressor)	0.722
SolarNN (Feedforward neural network)	0.717

Ten-fold cross-validation is carried out utilizing the census tract data. LR, linear regression; MARS, multivariate adaptive regression splines; RF, random forest. Hierarchical SolarForest proposed in the paper was the best-performing model.

Predictive Solar Deployment Models

We have developed and compared several non-linear machine learning models to estimate the census tract level solar deployment rate utilizing 88 environmental and socioeconomic factors as inputs. The models are linear regression with quadratic and interaction terms, multivariate adaptive regression splines (MARS), one-stage Random Forest, two-stage models utilizing a second stage with linear regression or MARS, two-stage Random Forest (SolarForest), and feedforward neural network (SolarNN). We utilize 10-fold cross-validation to estimate their out-of-sample performances. The results in Table 1 summarize performance utilizing cross-validation R^2 values (out-of-sample estimate) to compare easily between different models. SolarForest achieves R^2 = 0.722 and SolarNN achieves R^2 = 0.717, which are the highest state-of-the-art accuracy.

SolarForest is an ensemble Random Forest[29] framework with a Random Forest classifier and a Random Forest regression model (Figure S12). It aims at capturing a two-stage decision process at the census tract level. The classifier identifies whether a census tract has at least one system installed and the regressor estimates the number of systems installed in the tract in case the tract contains solar systems. Both models utilize the 88 socioeconomic and environmental census tract level variables listed in Supplemental Information Section 1.2. Gini importance is used to measure the feature importance for both classifier and regressor in the SolarForest, which is calculated by adding up the Gini impurity decreases during the fitting process for each individual feature. SolarNN is a feedforward neural network model with five hidden fully connected layers. Each hidden layer contains 88 neurons. It has a scalar output of the estimated value of solar deployment density. The activation function used in SolarNN is ReLU.[36]

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, 18 figures, and 4 tables and can be found with this article online at https://doi.org/10.1016/j.joule.2018.11.021.

ACKNOWLEDGMENTS

The authors would like to acknowledge Professor Susan Athey for discussions on building predictive adoption models. J.Y. thanks the support from State Grid of China for the Bits and Watts Fellowship. Z.W. thanks the support of the Stanford Interdisciplinary Graduate Fellowship (SIGF) as the Satre Family Fellow.

AUTHOR CONTRIBUTIONS

Conceptualization, J.Y., Z.W., A.M., and R.R.; Methodology, Z.W. and J.Y.; Software, Z.W. and J.Y.; Writing – Original Draft, J.Y., Z.W., A.M., and R.R.; Writing – Review & Editing, J.Y., Z.W., A.M., and R.R.; Funding Acquisition, A.M. and R.R.; Supervision, A.M. and R.R.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 6, 2018
Revised: August 22, 2018
Accepted: November 26, 2018
Published: December 19, 2018

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. Nat. Clim. 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. Appl. 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., and 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. Adv. Neural Inf. Process. Syst. 1097–1105.

12. Pan, S.J., and Yang, Q. (2010). A survey on transfer learning. IEEE Trans. Knowl. Data Eng. *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 Res. Soc. Sci. *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. Energ. Pol. *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. https://ases.conference-services.net/resources/252/2859/pres/SOLAR2012_0785_presentation.pdf.

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. Energ. Pol. *55*, 521–530.

19. Vasseur, V., and Kemp, R. (2015). The adoption of PV in the Netherlands: a statistical analysis of adoption factors. Renew. Sustain. Energ. Rev. *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 Res. Soc. Sci. *14*, 1–12.

21. Rai, V., Reeves, D.C., and Margolis, R. (2016). Overcoming barriers and uncertainties in the adoption of residential solar PV. Renew. Energ. *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 Res. Soc. Sci. *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. Energ. Pol. *109*, 141–153.

24. Davidson, C., Drury, E., Lopez, A., Elmore, R., and Margolis, R. (2014). Modeling photovoltaic diffusion: an analysis of geospatial datasets. Environ. Res. Lett. *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. Energ. Pol. *69*, 19–27.

27. De Groote, O., Pepermans, G., and Verboven, F. (2016). Heterogeneity in the adoption of photovoltaic systems in Flanders. Energ. Econ. *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 Res. Soc. Sci. *23*, 113–124.

29. Breiman, L. (2001). Random forests. Mach. Learn. *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. Sci. 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. Proc. IEEE Trans. Knowl. Data Eng. *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.