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

Jiafan Yu[*] **Zhecheng Wang**[*] **Arun Majumdar,** **Ram Rajagopal**
Stanford University

Abstract

We have developed DeepSolar, a deep learning framework that analyzes satellite imagery to identify the GPS locations and sizes of solar photovoltaic (PV) panels. Leveraging its high accuracy and scalability, we constructed a comprehensive high-fidelity solar deployment database for the contiguous U.S. We demonstrated its value by discovering that residential solar deployment density peaks at a population density of 1000 capita/mile^2, increases with annual household income asymptoting at $\sim$\$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 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.

1 Introduction

Deployment of solar photovoltaics (PV) is accelerating worldwide due to rapidly reducing costs and significant environmental benefits compared to electricity generation based on fossil fuels [Haegel et al., 2017]. Because of their decentralized and intermittent nature, cost-effective integration of solar panels on existing electricity grids is becoming increasingly challenging [Chu and Majumdar, 2012, Agnew and Dargusch, 2015]. 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 [NREL] 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 Furthermore, with the rapid pace of solar deployment, such a database could become outdated.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 [Jean et al., 2016]. The availability of satellite imagery with spatial resolution less than 30 cm for the majority of the U.S., which is annually updated, offers a rich data source for solar installation detection based on machine learning. Existing pixel-wise machine learning methods [Malof et al., 2016a, Yuan et al., 2016] suffer from poor computational efficiency, and relatively low precision and recall (cannot reach 85% simultaneously), while existing image-wise approach [Malof et al., 2016b] cannot provide system size or shape information. Google Inc.'s Project Sunroof utilizes a proprietary machine learning approach to report locations without any size information. They have so far identified much less number of systems (0.67 million) than in the Open PV database ($\sim$1 million) in the contiguous U.S.

[*]The first two authors contributed equally to this work.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

Leveraging the development of convolutional neural networks (CNNs) [LeCun et al., 2015] and large-scale labeled image datasets [Deng et al., 2009] for automatic image classification and semantic segmentation [Krizhevsky et al., 2012], 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 U.S. 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. 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. *More details can be found in full paper accepted by Joule Magazine.*

2 Results

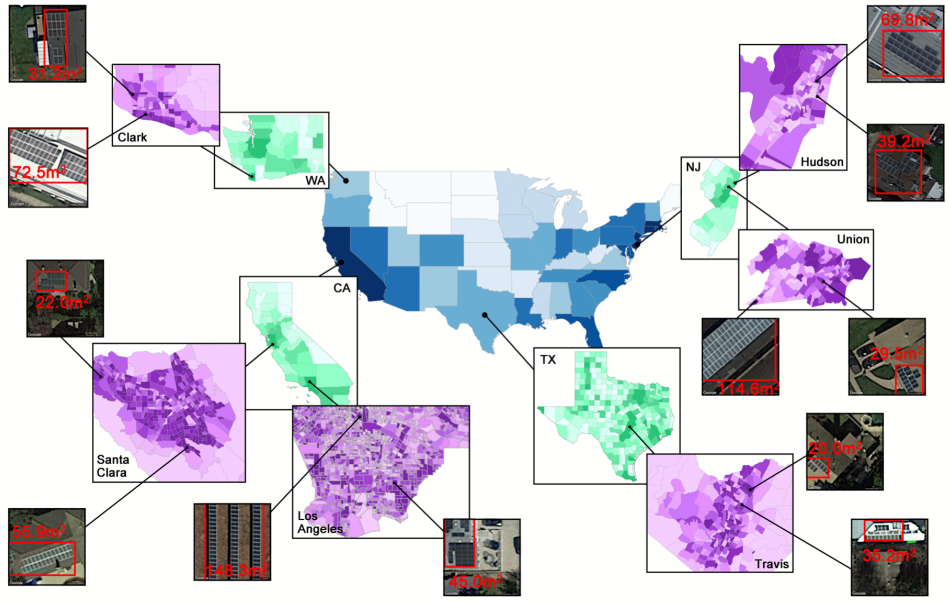
2.1 Scalable Deep Learning Model for Solar Panel Identification

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. Traditionally, training a CNN to 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 [Malof et al., 2016a, Yuan et al., 2016]. To enable efficient solar panel identification and segmentation, DeepSolar first utilizes transfer learning [Pan et al., 2010] to train a CNN classifier on 366,467 images sampled from over 50 cities/towns across the U.S. with merely image-level labels indicating 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 to 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 panel, greatly enhancing the computation efficiency.

The performance of our model is evaluated on a test set containing 93,500 randomly-sampled images across the U.S. 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 [Malof et al., 2016a, Yuan et al., 2016, Malof et al., 2016b,c]. 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.

2.2 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 U.S. with exact locations and sizes of solar panels. The number of detected solar systems in the contiguous U.S. is (1.4702$\pm$0.0007) million, which exceeds the 1.02 million installations without accurate location in Open PV [NREL] and the 0.67 million installations without



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 1: **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, NJ have solar resource density higher than 30,000 m^2/$mile^2$ while the five northern states (MT, ID, WY, ND, SD) 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.

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 solar farm. We built a complete resource density map in the contiguous U.S. from state level to household level (Fig. 1). Solar installation densities have dramatic variability at state (e.g., 1.34 to 224.1 m^2/$mile^2$) and county levels (e.g., 255 to 7490 m^2/$mile^2$ in CA). 23.4% of the census tracts contain 90% of the residential-scale installations.

2.3 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 U.S. 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 retail rate over the past 5 years, 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/d, which we define as an "activation" threshold triggering the increase of solar deployment. Since significant variation of solar deployment density is observed with solar radiation, we split all tracts into three groups according to the radiation levels (low, medium, high), and analyze the trends with other factors based on such grouping. Population/housing density has been observed to be positively [Schaffer and Brun, 2015] or negatively [Kwan, 2012, Crago et al., 2014] correlated with solar deployment. Fig. 2a shows that both trends hold but with a peak deployment density at the population density of 1000 capita/$mile^2$. Annual household income is a substantial driver for solar deployment (Fig. 2b). Low- and medium-income households have low deployment densities

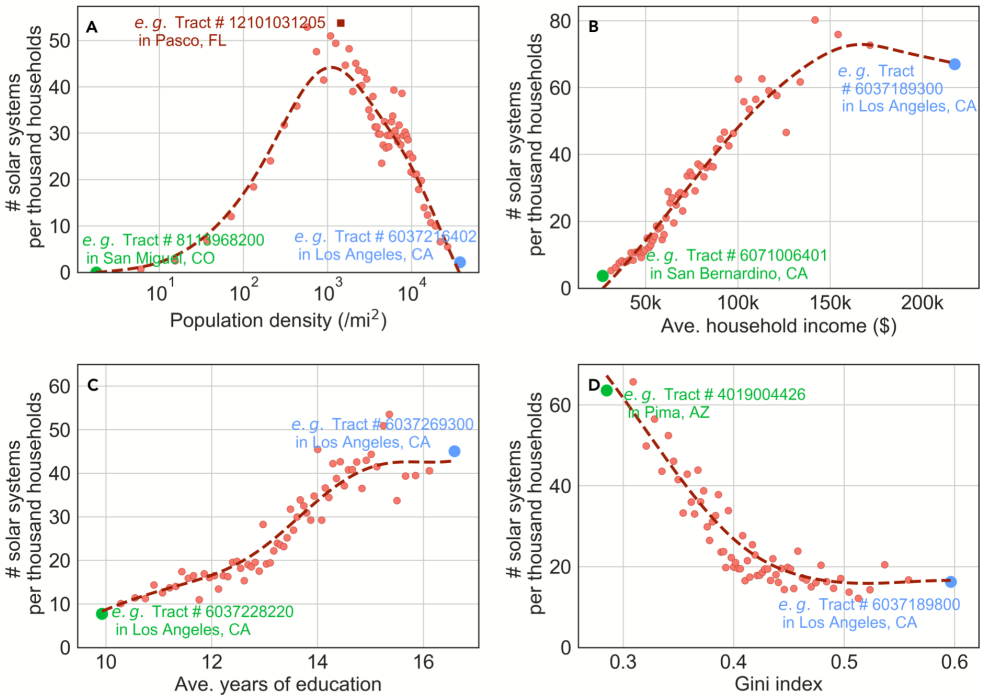


Figure 2: **Residential solar deployment density correlates with socioeconomic factors conditional on radiation** Census tracts are grouped according to 64 bins of the target factor. Curves are fitted utilizing LOWESS. Blue/green/brown label denotes the census tract and its affiliated county corresponding to the median value in the bin. Here we only show tracts with high solar radiation (>5.0 kWh/m^2/d). A. Solar deployment density increases with population density with a peak at 1000 capita/mi^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.

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 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 increasing trend with average education level (Fig. 2c). 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. Moreover, solar deployment density in census tracts with high radiation is strongly correlated and decreasing with the Gini index, a measure of income inequality (Fig. 2d). 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.

2.4 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 (e.g., [Vasseur and Kemp, 2015]) or data driven approaches (e.g., [De Groote et al., 2016]), achieving in-sample R^2 between 0.04 and 0.71. The models are typically linear or log-linear and utilize less than 10,000 samples for regression. Our result instead reveals that socioeconomic trends are highly nonlinear. Therefore, we build an Random-Forest-based model, called SolarForest, to estimate solar deployment at census tract level utilizing the data from more than 70,000 census tracts, which achieves the tier-1 out-of-sample R^2 of 0.72 in the ten-fold cross validation, higher than the in-sample R^2s of any other models in previous works.

References

Scott Agnew and Paul Dargusch. Effect of residential solar and storage on centralized electricity supply systems. *Nature Climate Change*, 5(4):315, 2015.

Steven Chu and Arun Majumdar. Opportunities and challenges for a sustainable energy future. *Nature*, 488(7411):294, 2012.

Christine Lasco Crago, Ilya Chernyakhovskiy, et al. 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*, pages 27–29, 2014.

Olivier De Groote, Guido Pepermans, and Frank Verboven. Heterogeneity in the adoption of photovoltaic systems in Flanders. *Energy Economics*, 59:45–57, 2016.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

Nancy M Haegel, Robert Margolis, Tonio Buonassisi, David Feldman, Armin Froitzheim, Raffi Garabedian, Martin Green, Stefan Glunz, Hans-Martin Henning, Burkhard Holder, et al. Terawatt-scale photovoltaics: Trajectories and challenges. *Science*, 356(6334):141–143, 2017.

Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

Calvin Lee Kwan. 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, 2012.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

Jordan M Malof, Kyle Bradbury, Leslie M Collins, and Richard G Newell. Automatic detection of solar photovoltaic arrays in high resolution aerial imagery. *Applied Energy*, 183:229–240, 2016a.

Jordan M Malof, Kyle Bradbury, Leslie M Collins, Richard G Newell, Alexander Serrano, Hetian Wu, and Sam Keene. Image features for pixel-wise detection of solar photovoltaic arrays in aerial imagery using a random forest classifier. *IEEE International Conference on Renewable Energy Research and Applications*, pages 799–803, 2016b.

Jordan M Malof, Leslie M Collins, Kyle Bradbury, and Richard G Newell. A deep convolutional neural network and a random forest classifier for solar photovoltaic array detection in aerial imagery, 2016c.

NREL. The Open PV Project. `https://openpv.nrel.gov`. Accessed: 2018-10-15.

Sinno Jialin Pan, Qiang Yang, et al. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

Axel J Schaffer and Sebastian Brun. Beyond the sun—socioeconomic drivers of the adoption of small-scale photovoltaic installations in Germany. *Energy Research & Social Science*, 10:220–227, 2015.

Véronique Vasseur and René Kemp. The adoption of PV in the Netherlands: A statistical analysis of adoption factors. *Renewable and Sustainable Energy Reviews*, 41:483–494, 2015.

Jiangye Yuan, Hsiu-Han Lexie Yang, Olufemi A Omitaomu, and Budhendra L Bhaduri. Large-scale solar panel mapping from aerial images using deep convolutional networks. *IEEE International Conference on Big Data*, pages 2703–2708, 2016.