An Investigation of Rooftop Solar in Philadelphia, PA

Anne Evered

CPLN505, Spring 2019

Contents

1 Introduction 2

2 Data and Methods 2

2.1 Dataset Overview 2

2.2 Data Preparation 2

2.3 Variables and Model Development 3

3 Results 4

3.1 Descriptive Statistics 4

3.2 Models for Percent of Buildings Qualified for Solar 5

3.3 Models for Percent of Qualified Buildings with Solar Installed 6

3.3.1 With Mean as Threshold 6

3.3.2 With Zero as Threshold 7

4 Discussion 9

4.1 Summary 9

4.2 Further Analyses 9

5 References 10

6 Appendix 11

1 Introduction

Rooftop solar panels provides one path for decreasing environmental degradation, reducing carbon emissions, and improving environmental health. In addition to the possible environmental benefits, the use of rooftop solar energy can also have economic benefits at both the household and community level. The costs of solar panel systems and installations have dropped significantly since first available and depending on specific roof space and sunlight conditions, conversion to solar electricity can lower overall utility costs.

Existing research, however, points to significant disparities in the adoption of solar technologies across both race and income categories. For example, a recent paper in *Nature Sustainability* found that for the same median household income, black- and Hispanic-majority census tracts had significantly less rooftop solar compared with no-majority tracts (by 69% and 30%, respectively), while white-majority census tracts had 21% more.[SCK19] Likewise, researchers have found significant differences in solar adoption across income categories, with rooftop solar adoption skewed toward higher income households prior to 2010 and shifting towards more moderate incomes since then.[Bar+18]

While this existing research provides a useful overview of national rooftop solar trends, individual cities deserve closer examination. In particular, a better understanding of city-specific trends in solar panel adoption can help ensure more equitable and sustainable energy policies within those cities. This analysis focuses specifically on the city of Philadelphia, PA and explores whether the spatial and demographic trends within the city mirror those found nationally. Findings from this analysis suggest that while race and income are both associated with higher levels of rooftop solar potential and adoption in Philadelphia, additional factors such as average household size and percent of households owner occupied may have greater effect sizes. The findings from this analysis also point to challenges in reliably and consistently measuring both rooftop solar potential and installations.

2 Data and Methods

2.1 Dataset Overview

The primary dataset for the analysis is Google's Project Sunroof, an open data portal that uses Google's data sources and mapping capabilities, along with data from the National Renewable Energy Laboratory (NREL), Aurora Solar, Clean Power Research and other sources to map solar use and potential in the United States. The portal allows users to search an address or geographic region (state, county, city, or zip code) and view estimates of solar potential and use based on the amount of usable sunlight and roof space (see Appendix for full list of variables). Along with the portal, the project also allows for export of the data at the state, county, city, postal code, and census tract level (as used in this analysis).

In addition to the Project Sunroof dataset, census tract level demographic and income measures were pulled from the 2006-2010 American Community Survey. Geographic information from OpenDataPhilly is used for all mapping.

2.2 Data Preparation

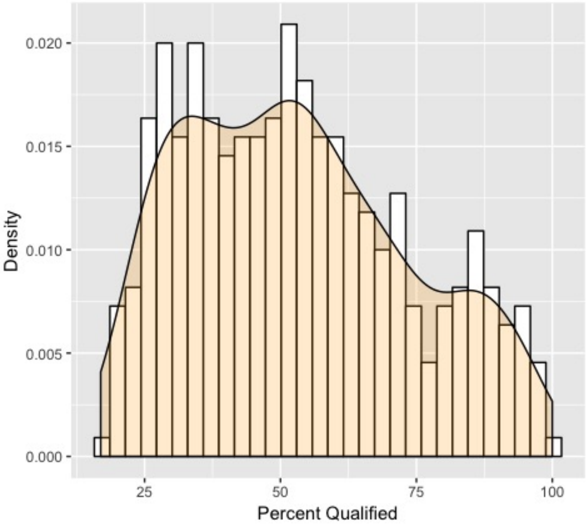
The datasets were joined on census tract ID and several data preparation were taken. Specifically, the original Project Sunroof dataset included 45 duplicate tracts (i.e. two rows per tract with different variable values). The Percent Covered (percent_covered) variable was used to determine which of the duplicate tracts to include in the dataset. Specifically, the Percent Covered metric indicates what percent of the census tract that is included in the data estimates for that tract. For the duplicate tracts, whichever row had a lower Percent Covered was dropped from the dataset. It was also confirmed that all tracts had at least 75 Percent Covered (minimum = 77.33%, mean = 93.02%). In addition, all tracts that had a population of zero based on the census data (8 tracts) were removed from the dataset. These data cleaning steps resulted in a final dataset of 375 rows at the tract level.

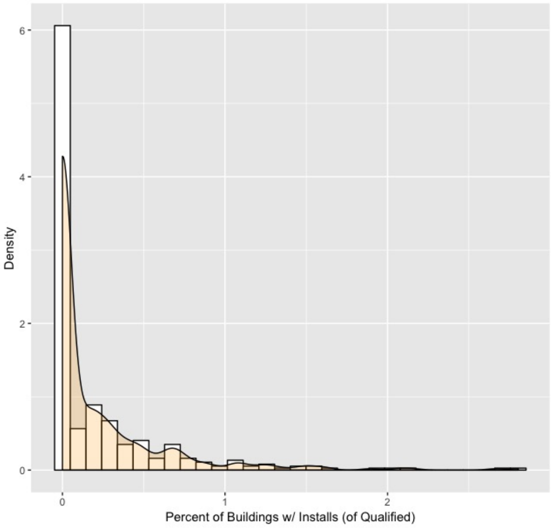
2.3 Variables and Model Development

To better understand rooftop solar in Philadelphia, two outcome variables were considered, one measuring rooftop solar potential and the other measuring rooftop solar installations:

1. Percent Qualified: the percent of buildings qualified for rooftop solar panels
2. Percent Installed: the percent of qualified buildings with rooftop solar panels installed

Percent Qualified is a pre-calculated metric of solar potential available in the Project Sunroof dataset. While public documentation does not define the exact calculation of this metric, it is likely based on the available sunlight and roof space. The second outcome variable of interest, Percent Installed, is a calculated metric derived from taking the number of buildings estimated to have rooftop solar within that census tract (count_qualified) divided by the total number buildings qualified for solar in that tract (existing_installs_count). The following histograms show the distributions for these two outcome variables:





The Percent Qualified metric is fairly normally distributed and therefore linear regression techniques are used to model this continuous quantity. In contrast, the Percent Installed variable has high degree of skew with over half of the tracts having a zero value. Therefore, to better model solar installations, a binary variable was created in place of the continuous installation variable and logistic regression techniques were used. Specifically, two thresholds are considered for the binary variable and presented in the results: 1) whether the Percent Installed is above or below the mean value (0.2121%) across all census tracts and 2) whether Percent Installed is greater than zero (i.e. the tract has at least one installation).

For both outcome variables, various demographic measures are considered as independent variables in the models. These variables are chosen primarily based on existing research and hypothesized factors related to solar use and potential, including:

- Median Age (years)
- Percent White
- Percent Hispanic or Latino
- Average Household Size
- Percent of Households with Income 0-$30,000

- Percent of Households with Income $30,000-$50,000
- Percent of Households with Income $50,000-$100,000
- Percent Owner-occupied

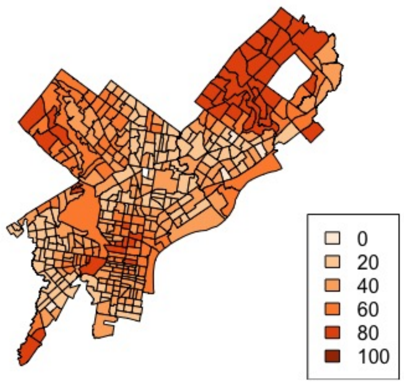
The Median Number of Panels per Building is also included to control for building roof size. Two additional variables, Percent Black and Percent of Households with Income over $100,000, are included in descriptive statistics, but removed from the models to avoid collinearity.

3 Results

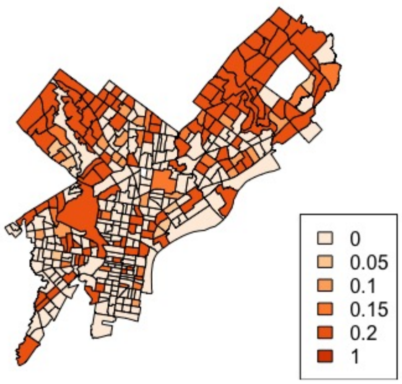
3.1 Descriptive Statistics

The following table and maps provide summary statistics and geographic distribution for the two outcome variables of interest. One important feature to note from the geographic distribution for Percent Qualified, is that based on the Project Sunroof dataset, there is a high skew in solar potential towards more suburban tracts. The assumption leading to this distribution should be explored in further analyses, as discussed later in the paper.

Variable	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Percent Qualified	16.93	35.44	50.39	52.37	66.07	97.24
Percent Installed	0.00	0.00	0.00	0.212	0.263	2.804



Percent of Buildings Qualified for Solar

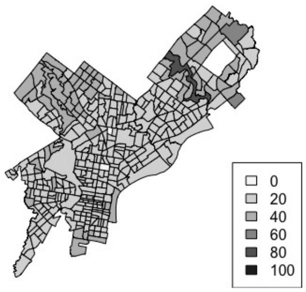


Percent of Buildings with Solar of Qualified

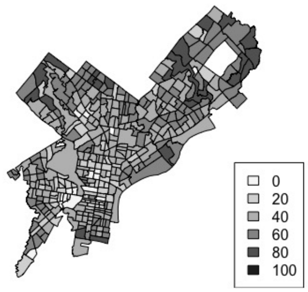
The following table provides summary statistics for the independent variables considered in the models. Maps of key variables of interest are also included to illustrate geographic distributions. Note that some tracts, for which the denominator of any of the percent variables was zero have been dropped from the dataset and therefore do not appear in the maps.

Variable	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Median Age (years)	19.70	30.65	34.00	35.06	38.55	82.40
Percent White	0.50	8.55	35.80	41.57	75.30	97.20
Percent Black	0.30	9.25	35.80	44.10	83.95	97.00
Percent Hispanic or Latino	1.00	2.80	4.30	10.86	9.10	91.20
Average HH Size	1.320	2.180	2.470	2.446	2.720	4.090
Population Density	1.351e-05	4.189e-03	6.929e-03	7.390e-03	9.773e-03	2.481e-02
Percent HH income <$30k	0.00	29.03	43.21	43.71	57.58	100.00
Percent HH income $30k-$50k	0.00	15.89	19.47	19.78	23.63	100.00
Percent HH income $50k-$100k	0.00	17.05	24.69	24.56	30.98	49.69
Percent HH income >$100k	0.00	3.962	8.257	11.950	16.481	68.137
Percent Owner-occupied	0.00	41.75	54.50	52.84	65.25	93.80
Median Number of Panels	10.00	12.00	14.00	24.21	20.00	854.00

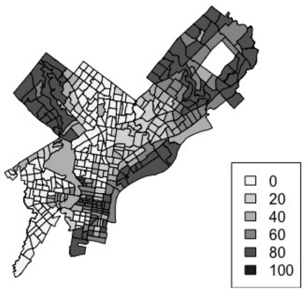
Median Age



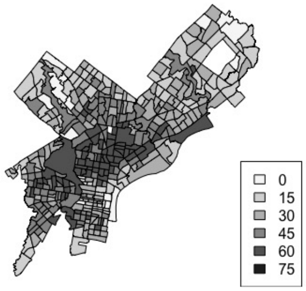
Percent Owner Occupied



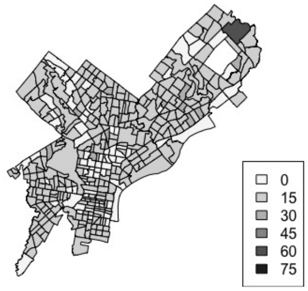
Percent White



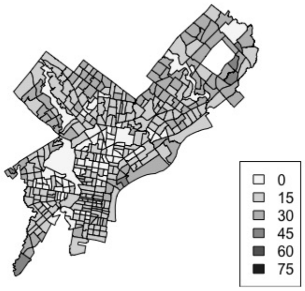
Percent HH Income <$30k



Percent HH Income $30k-50k



Percent HH Income $50k-100k



3.2 Models for Percent of Buildings Qualified for Solar

The initial model run for Percent Qualified is a linear regression with the ten independent variables discussed above. The results, with variables standardized around the mean, of this initial model can be seen in the below table:

Variable	Coefficient	95% Confidence Interval	P-value	sqrt (VIF)
(Intercept)	52.374	(51.163, 53.585)	< 2e-16	-
Median Age (years)	7.0502	(5.314, 8.787)	1.88e-14	1.431
Percent White	8.0763	(6.371, 9.782)	< 2e-16	1.406
Percent Hispanic or Latino	-2.1222	(-3.759, -0.485)	0.01120	1.350
Average Household Size	3.5253	(1.164, 5.886)	0.00353	1.947
Percent Owner Occupied	-13.8854	(-16.170, -11.600)	< 2e-16	1.884
Population Density	-6.1219	(-7.446, -4.798)	< 2e-16	1.0921
Percent HH Income $30k-$50k	-1.4724	(-2.927, -0.0183)	0.04720	1.199
Percent HH Income $50k-$100k	3.2168	(1.618, 4.815)	9.13e-05	1.318
Percent HH Income >$100k	2.4412	(0.636, 4.247)	0.00818	1.489
Median Number of Panels	2.0783	(0.686, 3.470)	0.00353	1.148

This initial model has an adjusted R-squared value of 0.6512. To assess for any collinearity, the square root of the variance inflation factor was taken and all coefficients were confirmed to have a value < 2 (values can be found in last column of above table). The following variables were significant at a level of < .01: Median Age, Percent White, Percent Owner Occupied, Population Density, and Percent HH Income $50k-$100k. Based on this preliminary model, a second simpler model was also run with just these most significant variables. The results of this model are below:

Variable	Coefficient	95% Confidence Interval	P-value	sqrt (VIF)
(Intercept)	52.374	(51.128, 53.620)	< 2e-16	-
Median Age (years)	6.811	(5.335, 8.287)	< 2e-16	1.183
Percent White	8.606	(7.207, 10.004)	< 2e-16	1.121
Percent Owner-Occupied	-12.594	(-14.115, -11.073)	< 2e-16	1.219
Population Density	-6.632	(-7.969, -5.295)	< 2e-16	1.0718
Percent HH Income $50k-$100k	2.973	(1.447, 4.499)	0.00015	1.223

This revised, simplified model has an adjusted R-squared value of 0.6307, only a slight reduction from the more complex model. An ANOVA test for the two models presented above was also run, giving a p-value of 3.227e-07.

The coefficients in the simplified model suggest the following: A one standard deviation increase in age is associated with a 6.811 increase in the percent of buildings suitable for solar (Percent Qualified). A one standard deviation increase in Percent White is associated with a 8.606 increase in the Percent Qualified. A one standard deviation increase in the Percent Owner Occupied is associated with a 12.594 decrease in the Percent Qualified. A one standard deviation increase in the Population Density is associated with a 6.632 decrease in the Percent Qualified. A one standard deviation increase in Percent HH Income $50k-$100k is associated with a 2.973 increase in the Percent Qualified.

3.3 Models for Percent of Qualified Buildings with Solar Installed

As discussed in the methods section, for modeling solar installations through the Percent Installed variable, two binary variables with different thresholds (the mean and zero) are considered. The results of models for both of these variables are presented below.

3.3.1 With Mean as Threshold

For the first set of Percent Installed models, the mean (0.2121 %) is used as the binary variable threshold. This threshold gives a breakdown of 264 tracts below the mean and 111 tracts above the mean. As with models for Percent Qualified, an initial model with all ten independent models standardized about the mean was run. The results of this model (on the odds ratio scale) can be seen in the following table.

Variable	Odds Ratio	95% Confidence Interval	P-value
(Intercept)	0.3297	(0.2463, 0.4327)	9.56e-15
Median Age (years)	1.6411	(1.1374, 2.4459)	0.01098
Percent White	1.4389	(1.0131, 2.0554)	0.04308
Percent Hispanic or Latino	0.6856	(0.4236, 1.0314)	0.09175
Average Household Size	2.3465	(1.3673 4.1846)	0.00275
Percent Owner-Occupied	0.4487	(0.2675, 0.7328)	0.00177
Population Density	0.4196	(0.2911, 0.5868)	1.13e-06
Percent HH Income $30k-$50k	1.1600	(0.8424, 1.6289)	0.37668
Percent HH Income $50k-$100k	1.2331	(0.8791, 1.7315)	0.22381
Percent HH Income >$100k	1.8765	(1.2717,2.8362)	0.00205
Median Number of Panels	0.8447	(0.5110, 1.2233)	0.50855

The AIC for this model is 381.84 and the McFadden's R-squared value is 0.210. The following variables were significant at a level of < .01: Average Household Size, Percent Owner Occupied, Population Density, and Percent HH Income >$100k.

Based on this preliminary model, a second simpler model was also run with just the most significant variables. The results of this model are below:

Variable	Odds Ratio	95% Confidence Interval	P-value
(Intercept)	0.3391	(0.2570, 0.4402)	2.90e-15
Average Household Size	1.1809	(0.8217, 1.7025)	0.369165
Percent Owner-Occupied	0.8921	(0.6454, 1.2307)	0.486747
Population Density	0.3621	(0.2572, 0.4955)	1.17e-09
Percent HH Income >$100k	1.9201	(1.3691, 2.7457)	0.000232

The new AIC for this model is 386.69 (compared to 381.84 for the more complex model) and the McFadden's R-squared value is 0.173. This model suggests that a one unit increase in Population Density is associated with 63.79% decrease in the odds of the Percent Installs for a given tract being above the mean and a one unit increase in the Percent HH Income >$100k is associated with a 92.01% increase in the odds of the Percent Installs being above the mean.

3.3.2 With Zero as Threshold

In addition to the above mean threshold, a threshold of zero (i.e. whether the tract has at least one building with solar installations) is also modeled. The zero valued threshold has the advantage of greater ease of interpretation compared to the mean threshold. The breakdown for this variable is 220 tracts at zero and 155 tracts greater than zero.

Again, an initial model with all ten independent variables standardized about the mean was run. The results of this model (on the odds ratio scale) can be seen below.

Variable	Odds Ratio	95% Confidence Interval	P-value
(Intercept)	0.6614	(0.5203, 0.8370)	0.000637
Median Age (years)	1.4719	(1.0381, 2.1486)	0.036504
Percent White	1.6015	(1.1567, 2.2335)	0.004905
Percent Hispanic or Latino	0.7782	(0.5481, 1.0754)	0.141232
Average Household Size	2.1098	(1.3073, 3.5027)	0.002877
Percent Owner-Occupied	0.5290	(0.3300, 0.8292)	0.006583
Population Density	0.4756	(0.3515, 0.6306)	5.93e-07
Percent HH Income $30k-$50k	0.4795	(0.2611, 0.8551)	0.014926
Percent HH Income $50k-$100k	0.8123	(0.5803, 1.1270)	0.218291
Percent HH Income >$100k	0.8693	(0.5561, 1.3373)	0.530048
Median Number of Panels	0.7692	(0.5051,1.1103)	0.213934

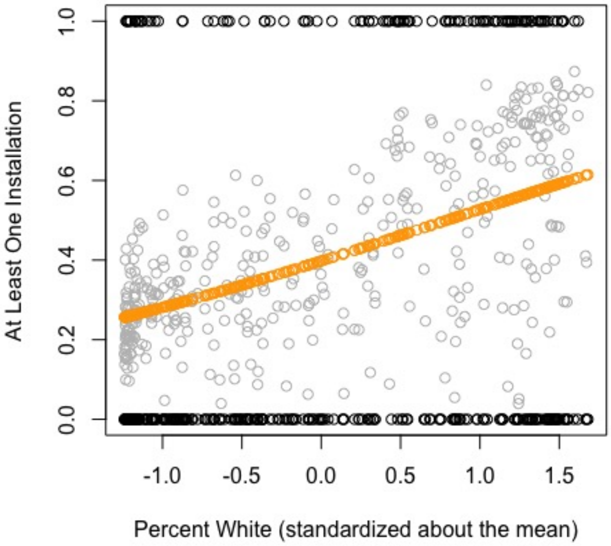
The AIC for this model is 440.92 and the McFadden's R-squared value is 0.1762. The following variables were significant at a level of < .01: Percent White, Average Household Size, Percent Owner Occupied, and Population Density. In contrast to the mean threshold outcome variable, in this case, none of the income variables were significant at a level < .01 (though Percent HH Income $30k-$50k is significant at level 0.014926).

Based on this preliminary model, a second simpler model was also run with just the most significant variables. The results of this model are below:

Variable	Odds Ratio	95% Confidence Interval	P-value
(Intercept)	0.6603	(0.5234, 0.8292)	0.000398
Percent White	1.6893	(1.3036506 2.2069528)	9.15e-05
Average Household Size	1.0308	(0.7750663 1.3705774)	0.834284
Percent Owner-Occupied	1.0632	(0.8111961 1.3949273)	0.656500
Population Density	0.4539	(0.3418790 0.5904019)	1.37e-08

The new AIC for this model is 446.04 (compared to 440.92 for the more complex model) and the McFadden's R-squared value is 0.1426. These model results suggest that a one standard deviation increase in Percent White is associated with a 68.93% increase in the odds of having at least one solar installation. A one standard deviation increase in the Population Density is associated with a 0.5461% decrease in the odds of having at least one solar installation.

In contrast to the mean threshold model, Percent White has a relatively high effect size and significance in this model. The following plot specifically focuses on this variable and illustrates the the difference in Percent White from the mean against the observed, fitted, and predicted values (with all other variables held at their mean value) for whether there is at least one installation (Percent Installs > 0).



In summary, the results of the models for the binary Percent Installs variable depend on the particular threshold set. The models for both of the thresholds considered, though, suggest that while the income and race categories do have some effect on the odds of there being a certain percentage of installs, other variables such as Average Household Size and Percent Owner-Occupied appear to have a more consistent effect, at least based on this solar dataset.

4 Discussion

4.1 Summary

In conclusion, this analysis finds some significant associations between the racial and economic characteristics of Philadelphia census tracts and the suitability for and adoption of rooftop solar within those tracts. These findings mirror national trends from the existing research discussed in the introduction of this paper. Specifically, based on initial models using the Project Sunroof dataset, both the percent of buildings qualified for solar and the percent of buildings with solar installed appear to be associated with higher percentages of higher income levels and white populations.

Areas with high solar potential and solar installations, though, also appear to be heavily skewed towards more suburban, less-dense districts in the Project Sunroof dataset. The validity of this aspect of the Project sunroof data deserves closer examination, particularly because this feature may account for some of the findings discussed above.

Moreover, despite these findings, in none of the models was the effect size for the race and income variables greatest among those included in the models. For example, the Percent of Households Owner Occupied and Population Density both had a larger negative effect on Percent Qualified and Average Household size had a greater positive effect on Percent Installed.

4.2 Further Analyses

There are several areas for further investigation from this analysis. First, further analysis should investigate the assumptions of the rolled-up metrics in the Project Sunroof dataset. As discussed throughout the paper, the Project Sunroof dataset suggests significantly higher solar potential in suburban, less-dense tracts. This calls into question what assumptions are used in the calculations of the metrics. Additional solar data, or the raw data used to create the Project Sunroof dataset, could be used to better understand this aspect of the dataset and, in general, to validate the assumptions made in the rolled-up metrics.

The Project Sunroof dataset also does not provide information on change over time. Further analysis could supplement the dataset with other solar data to investigate changes in solar adoption over time and to better understand the impact of policy decisions on these metrics. In addition, other independent variables, such as education level or job density, could be included in future investigations. This initial analysis also does not separate between commercial and residential buildings, a distinction that could be incorporated in future models given that the census demographic information only applies to residential buildings.

Lastly, comparing solar trends in Philadelphia to those in other like-cities, particularly cities with similar renewable energy goals, would be a good next step for this analysis. Gaining a better understanding of solar adoption trends in Philadelphia compared to like-cities could further help to inform where additional resources may be needed for residential solar panel adoption.

5 References

[Bar+18] Galen Barbose et al. “Income Trends of Residential PV Adopters”. In: *Lawrence Berkeley National Laboratory* (2018). DOI: https://www.cesa.org/assets/2018-Files/income-trends-of-residential-pv-adopters.pdf.

[SCK19] Deborah A. Sunter, Sergio Castellanos, and Daniel M. Kammen. “Disparities in rooftop photovoltaics deployment in the United States by race and ethnicity”. In: *Nature Sustainability* 2 (2019), pp. 71–76. DOI: https://www.nature.com/articles/s41893-018-0204-z.

6 Appendix

List of Project Sunroof Variables

Variable	Description
region_name	Region name (census tract, zip code, city, county, state)
count_qualified	of buildings in Google Maps that are suitable for solar
percent_covered	% of buildings in Google Maps covered by Project Sunroof
percent_qualified	% of buildings covered by Project Sunroof that are suitable for solar
yearly_sunlight_kwh_kw_threshold_avg	75% of the optimum sunlight in the county containing that zip code
yearly_sunlight_kwh_total	total solar energy generation potential for all roof space in that region
yearly_sunlight_kwh_f	total solar energy generation potential for flat roof space in that region
yearly_sunlight_kwh_n	total solar energy generation potential for north-facing roof space in that region
yearly_sunlight_kwh_e	total solar energy generation potential for east-facing roof space in that region
yearly_sunlight_kwh_s	total solar energy generation potential for south-facing roof space in that region
yearly_sunlight_kwh_w	total solar energy generation potential for west-facing roof space in that region
number_of_panels_total	of solar panels potential for all roof space in that region, assuming 1.650m x 0.992m panels
number_of_panels_f	of solar panels potential for flat roof space in that region, assuming 1.650m x 0.992m panels
number_of_panels_n	of solar panels potential for north-facing roof space in that region, assuming 1.650m x 0.992m panels
number_of_panels_e	of solar panels potential for east-facing roof space in that region, assuming 1.650m x 0.992m panels
number_of_panels_s	of solar panels potential for south-facing roof space in that region, assuming 1.650m x 0.992m panels
number_of_panels_w	of solar panels potential for west-facing roof space in that region, assuming 1.650m x 0.992m panels

Variable	Description
carbon_offset_metric_tons	The potential carbon dioxide abatement of the solar capacity that meets the technical potential criteria. The calculation uses eGRID subregion CO2 equivalent non-baseload output emission rates. https://www.epa.gov/sites/production/files/2015-10/documents/egrid2012_summarytables_0.pdf
number_of_panels_median	of panels that fit on the median roof
yearly_sunlight_kwh_median	kWh/kw/yr for the median roof, in DC (not AC) terms
install_size_kw_buckets_json	of buildings with potential for various installation size buckets. Format is a JSON array, where each element is a tuple containing (1) lower bound of bucket, in kW, and (2) number of buildings in that bucket.
lat_avg	average latitude for that region
lng_avg	average longitude for that region
lat_min	minimum latitude for that region
lng_min	minimum longitude for that region
lat_max	maximum latitude for that region
lng_max	maximum longitude for that region
state_name	Name of the state containing that region
kw_total	of kW of solar potential for all roof types in that region (assuming 250 watts per panel)
kw_median	kW of solar potential for the median building in that region (assuming 250 watts per panel)
existing_installs_count	of buildings estimated to have a solar installation, at time of data collection