Equitability of Solar Energy in the United States: Analysis of Factors Governing Urban Photovoltaic System Distribution

By Marshall Roll

Abstract

As American cities aim to cut carbon emissions, the broad uptake of photovoltaic systems (PVs) is crucial. However, previous research (Cook and Bird 2018; Barbose et al. 2021) suggests that PV distribution in American cities may be inequitable, which is harmful from both an environmental justice and carbon footprint mitigation perspective. This study aims to determine if PV distribution is proceeding inequitably by examining social and spatial trends governing installation levels in census tracts across several major US cities using linear and LASSO regression modeling. We find that there are significant social and spatial trends that determine PV installation rates in San Diego, but not in other studied cities. This modeling helps to determine which communities could benefit the most from future PV installation and provides insight into understanding how cities can distribute renewable energy more equitably.

Introduction

Solar energy is one of the most burgeoning renewable energy sources, quickly increasing in both efficacy and availability (Barbose et al. 2021). Photovoltaic (PV)[1] retrofitting is an increasingly common process for households and businesses alike, as it helps to decrease carbon footprints and cut long-term energy costs (IRENA 2021). This is crucial both from an economic and climate change mitigation perspective, as largescale shifts to solar energy would help to reduce fossil fuel combustion while also potentially lowering energy costs for consumers. (IRENA 2021; US Department of Energy 2021). Although there are still issues associated with solar energy, such as the scarcity of silicon and the necessity for development of energy storage capabilities (Zsiborács et al., 2019), it is an accessible and important renewable energy option (Jakhongir, 2019; US Department of Energy 2021).

Throughout the 2010s, PVs have primarily been implemented by businesses and high-income families while remaining inaccessible to many low-income families (O'Shaughnessy et al. 1). Although there are solar incentive tax credits available to potential consumers, PV systems have remained inaccessible for many low and middle-income families, who often rent their homes or live in apartments and thus have no incentive to invest in such infrastructure (Office of Energy Efficiency & Renewable Energy 2021; Cook and Bird 2018). Also, many low-income families live in houses built before 1970, which can provide an additional barrier for PV purchase due to outdated electrical infrastructure (Cook and Bird 2018). However, recent research suggests that PV uptake may be increasing in low-income families above the poverty line, providing an update on the United States' ability to achieve its energy goals (Barbose et al. 2021).

This research will continue in that vein by focusing on the trends governing current urban solar usage. We aim to determine which predictors are the most significant in determining PV system distribution in urban areas in the United States. We suspect that geographic patterns, such as annual sunlight levels, affect PV system location. Based on previous research, we also

[1] We define photovoltaic systems as one or more working modules capable of transforming solar energy into a usable electrical form. (Solar Energies Technology Office)

hypothesize that social factors such as race and median household income will be significant factors impacting PV system distribution. (Cook and Bird 2018; Barbose et al. 2021)

Methods

Area of Study

To compare trends of equitable photovoltaic system (PV) distribution across the United States, four cities of study were selected based on their population, PV installation rate, and solar radiation levels. We restricted our area of study to counties containing metropolitan areas, which are defined as urban settings with more than 500,000 inhabitants. (OECD) Analysis is conducted in all census tracts within a county to provide an understanding of equitable solar distribution along the urban gradient and to facilitate simpler data processing.

Chicago was chosen to represent a city with low installation and low solar radiation, Portland was chosen as a city with high installation and low solar radiation, San Diego was selected as a city with high installation and high solar radiation, and Houston was selected with low installation and high solar radiation. (Table 1) In this way, the analysis encompassed cities in every combination of solar radiation and existing installation levels. Data was then extracted from Google Sunroof, Stanford University's DeepSolar project, and 2020 Census Tract Data for the counties in which these cities are located.

County	Cook (Chicago)	Multnomah (Portland)	San Diego (San Diego)	Harris (Houston)
Population	5,173,146	803,377	3,286,069	4,728,030
Average solar radiation (kWh/yr)	973	850	1,294	1,063
Rate of PV installation (solar systems/house)	0.97	11.70	88.54	1.03

Table 1: Qualifying criteria of the case studies, showing metropolitan area population, average annual solar radiation levels, and the rate of PV system installation, measured in terms of the number of solar systems per house.
Source: Census, 2021; Google Sunroof; DeepSolar project, 2018

Data Analysis

Analysis was conducted in R using two techniques. First, we computed our own linear regression model using the variables in our hypothesis to test whether income, solar radiation, and race are statistically significant determinants of PV installation. This equation takes the form:

$$solar\_installation = \beta_1 * income + \beta_2 * radiation + \beta_3 * race$$

where β values are constants, *income* is median household income for a census tract, *radiation* is the average solar radiation in kWh/year, and *race* is the proportion of nonwhite people per census tract.

Second, we performed a LASSO (Least Absolute Shrinkage and Selection Operator) variable selection using the variables listed in Table 2.[2] This algorithm ranked the variables in order of their importance as predictors of PV system installation and calculated the coefficients of the selected variables in a regression model. The predictor variables inputted to the model exhibit some covariance and broadly fall into two categories: social and spatial predictors. Social predictors, such as income, race, educational attainment, rate of residents using a car as their sole mode of transportation, and median age are all related factors that may affect the distribution of PV installation. Spatial factors, such as yearly sunlight levels and predicted carbon offset, affect the practicality of PV installation—areas with less sunlight are less likely to benefit from solar systems. If PV installation is equitable, spatial factors will be the most significant predictors, *ceteris paribus*.

- Average and Median Household Income
- Percentage of Families below Poverty Level
- Rate of Educational Attainment
- Race
- Electricity Price
- Electricity Demand
- Housing Value
- Median Age
- Age Proportion
- Car Owner Percentage
- Rate of Residents Using Car as Sole Transportation
- Public Health Insurance Rate
- Gini Index
- Political Affiliation
- Presence of Fiscal Incentives
- Yearly Sunlight
- Predicted Carbon Offset

Table 2: Variables input to LASSO
Source: Census, 2021; Google Sunroof; DeepSolar project, 2018

Results

Preliminary analysis shows that PV distribution is greatly right-skewed in all four cities. The maximum PV area per capita in a single selected census tract is 4.388 m^2, indicating that there is a high level of variability. San Diego has the largest degree of variability, with PV area per capita distributed non-normally. (Fig. 1)

[2] The specific algorithm used was the "glmnet" method of the "train" function in R's "caret" package.

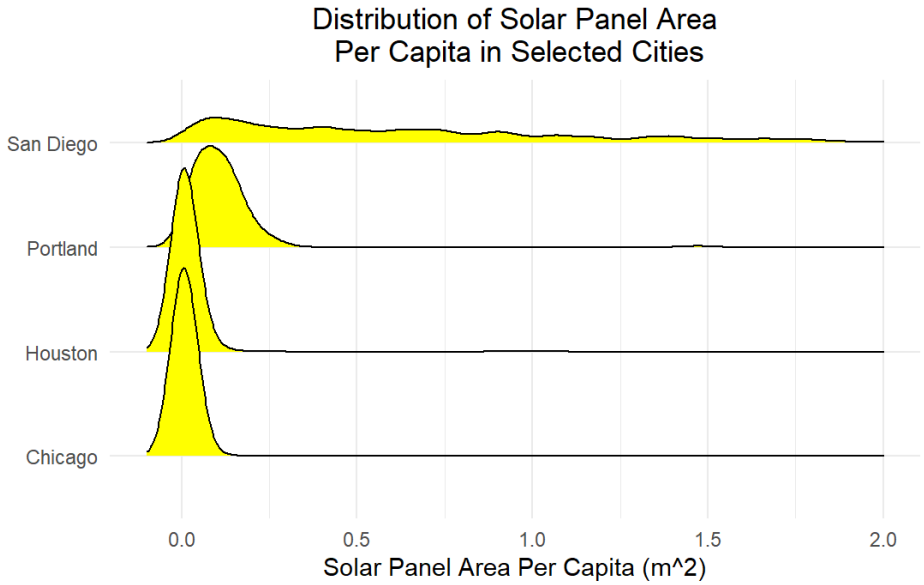

Figure 1: The distribution of PV area per capita (m^2) in each census tract of selected counties. Census tract sample sizes are $N = 619$ in San Diego, $N = 181$ in Portland, $N = 813$ in Houston, and $N = 1350$ in Chicago.

Further analysis shows that San Diego's PV area per capita exhibits the greatest amount of variation with changes in the selected predictors. San Diego's PV area per capita tends to increase with increasing median household income (Fig. 2), slightly decrease with increasing percentages of nonwhite population (Fig. 3), and increase with increasing sunlight levels (Fig. 4). Portland exemplify the same general trends, though much more weakly. Chicago and Houston do not have significant relationships with the predictor variables as seen in Figures 2, 3, and 4.

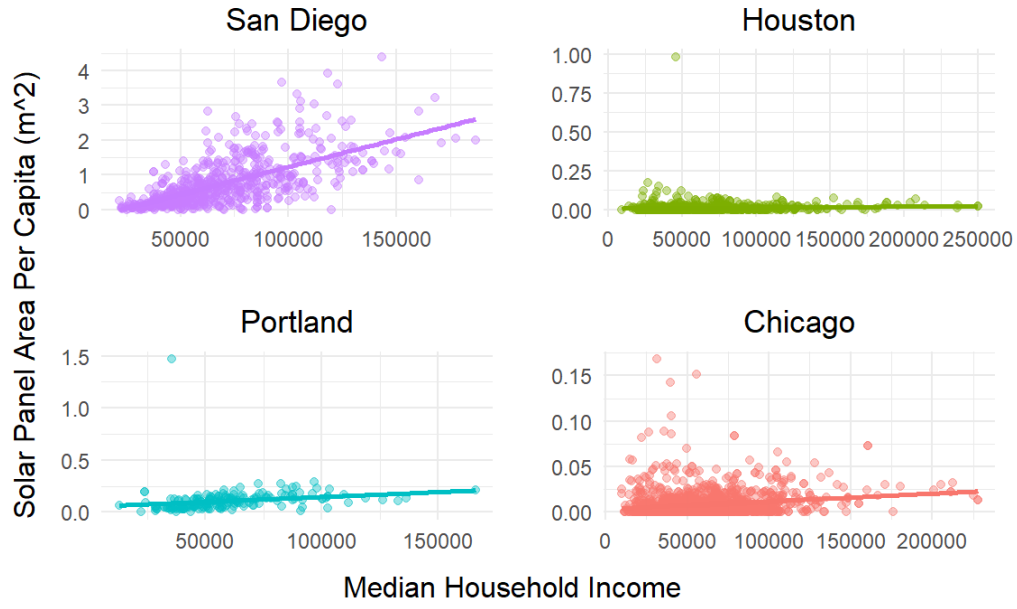

Figure 2: The relationship between solar panel area per capita and median household income for census tracts in each city (*N* = 619 in San Diego, *N* = 181 in Portland, *N* = 813 in Houston, and *N* = 1350 in Chicago). Trendlines show line of best fit as predicted by linear regression modeling.

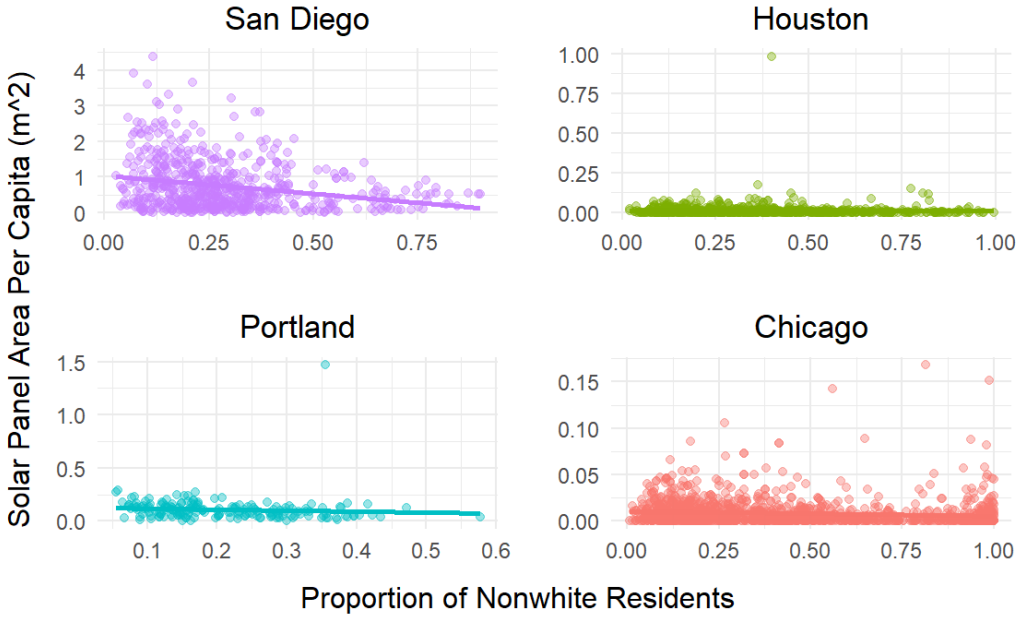

Figure 3: The relationship between solar panel area per capita and the proportion of nonwhite residents for census tracts in each city (*N* = 619 in San Diego, *N* = 181 in Portland, *N* = 813 in Houston, and *N* = 1350 in Chicago). Trendlines show line of best fit as predicted by linear regression modeling.

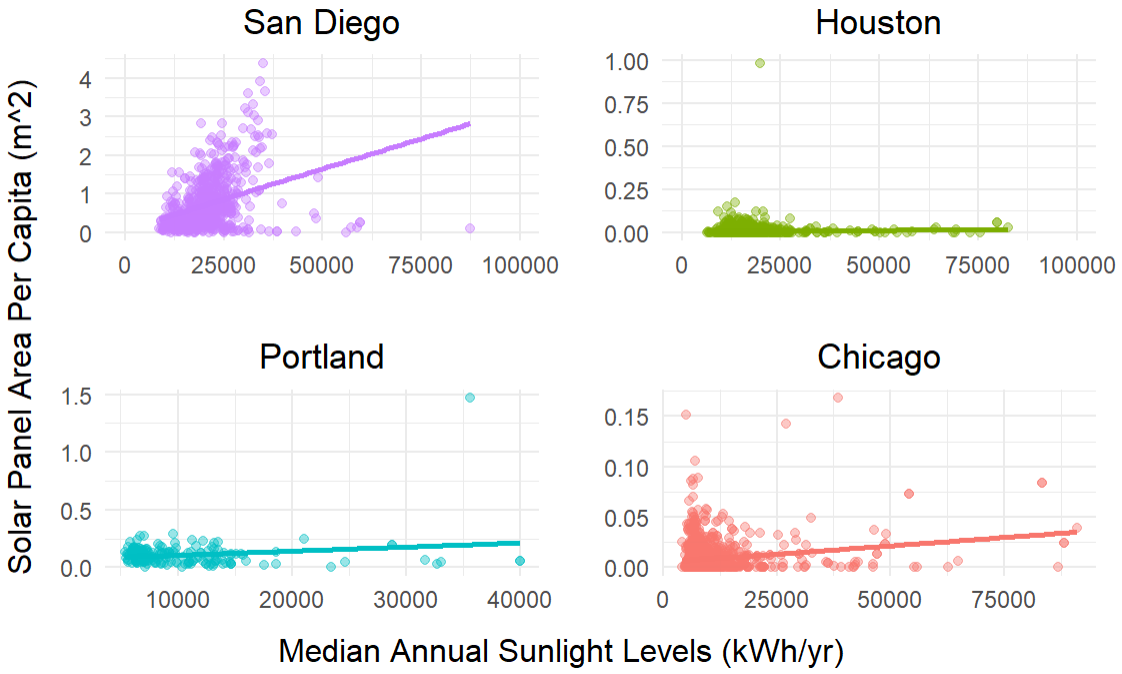

Figure 4: The relationship between solar panel area per capita and the median annual sunlight levels for census tracts in each city (*N* = 619 in San Diego, *N* = 181 in Portland, *N* = 813 in Houston, and *N* = 1350 in Chicago). Trendlines show line of best fit as predicted by linear regression modeling.

Simple Regression

The simple linear regression models included three variables that we hypothesized were likely predictors of solar panel area per capita: race, income, and sunlight. For Chicago, Houston, and Portland, models based on these three variables hold weak predictive power. However, the regression model for San Diego County is reasonably strong, with all variables exhibiting statistical significance. (Table 3)

City	Linear Regression	R^2
San Diego	Solar panel area per capita = -2.60×10^{-1}** $-5.74 \times 10^{-1} \cdot$ *race* *** $+1.50 \times 10^{-5} \cdot$ *income* *** $+6.25 \times 10^{-6} \cdot$ *sunlight* ***	0.463
Portland	Solar panel area per capita = -4.45×10^{-2} $+7.42 \times 10^{-2} \cdot$ *race* $+1.44 \times 10^{-6} \cdot$ *income* ** $+4.95 \times 10^{-1} \cdot$ *sunlight* **	0.004
Houston	Solar panel area per capita = $+6.44 \times 10^{-3}$ $-4.33 \times 10^{-5} \cdot$ *race* $+6.84 \times 10^{-8} \cdot$ *income* $+7.54 \times 10^{-8} \cdot$ *sunlight*	0.004
Chicago	Solar panel area per capita = $+5.51 \times 10^{-4}$ $+3.56 \times 10^{-4} \cdot$ *race* $+7.88 \times 10^{-8} \cdot$ *income* *** $+3.03 \times 10^{-7} \cdot$ *sunlight* ***	0.063
		*p<0.05 **p<0.005 ***p<0.0005

Table 3: Results of the regression analysis for each city with solar panel area per capita as the outcome variable and *race* (proportion of nonwhite residents in a census tract), *income* (median household income in a census tract), and *sunlight* (median annual sunlight level in kWh) as predictor variables.

LASSO Regression

To conduct a more thorough analysis of which factors governed equitable solar panel distribution, we used a LASSO regression model. This machine learning algorithm automatically determines which predictor variables are statistically significant in predicting the outcome variable—in this case the solar panel area per capita for a census tract.

For each city, we set the training method to *cv*, training number to 10, and selection function to *oneSE*. Instead of outputting the "best" model with the lowest mean absolute error (MAE), the *oneSE* selection function outputs the model with the fewest number of predictors that has an MAE within one standard error of the lowest MAE. For this context, *oneSE* is the best selection function because the resulting model yields interpretations with real-world applications, rather than one for black box calculations. Lastly, we selected the lambda value range through trial and error. These lambda ranges were chosen with the goal of finding a "goldilocks" phenomenon

with the MAE, where the model is the proper degree of complexity without overfitting. For San Diego, Portland, Houston, and Chicago, our lambda ranges were set between 10^{-3} and $10^{-1.5}$, 10^{-3} and 10^{0}, 10^{-3} and 10^{-1}, and 10^{-3} and 10^{-2}, respectively.

San Diego's LASSO (Table 4) selected the following variables (+: positive correlation, −: negative correlation) as significant predictors in a linear regression for solar panel area per capita: average household income (+), median household income (+), rate of getting education higher than high school (−), rate nonwhite (−), median age (+), the rate of using a car as sole mode of transportation (+), the yearly median sunlight in kWh (+), and the hypothetical tons of carbon offset by solar panels (+). The R^2 of this model is 0.486, indicating that it is not very strong, as slightly more than half of the data's variance is left unexplained. However, given the broad context of the data, this R^2 may be deemed as acceptable. The final model's MAE is 0.347, which is also acceptable because the solar area per capita ranges from 0 to 4.3—which is large given the context of the data. Lastly, to determine whether this model is accurate, we examined its residuals, the distance between actual and predicted solar panel area per capita for a given census tract. Residuals are generally balanced above and below the x-axis (Fig. 5), except in census tracts where there was a low predicted solar panel area per capita, where they tend to be more positive. This is indicative of a well-fitting model except for near low predicted solar panel area per capita, where the model systemically overpredicts. The residuals also increase in magnitude as the predicted solar panel area per capita increases indicating a degree of heteroskedasticity. This means that the model becomes less reliable as solar panel area per capita increases.

San Diego LASSO Results			
Predictor	**Coefficient**	**Model Statistics**	
Intercept	−1.50	**R^2**	0.486
Average Household Income	8.49×10^{-7}	**MAE**	0.347
Median Household Income	9.74×10^{-6}		
Proportion of Residents with Less than High School Education	-9.90×10^{-2}		
Proportion of Nonwhite Residents	-1.69×10^{-1}		
Median Age	2.09×10^{-2}		
Proportion of Residents Using Only a Car to Drive to Work	9.41×10^{-1}		
Median Yearly Sunlight (kWh)	2.53×10^{-6}		
Potential Carbon Offset (metric tons)	7.25×10^{-6}		

Table 4: Coefficients for the LASSO regression model for San Diego County based on data inputted from $N = 619$ census tracts. Positive coefficients indicate that the predictor has a positive relationship with solar panel area per capita and negative coefficients show that the predictor has a negative relationship with solar panel area per capita. R^2 and mean absolute error of the model are reported in the model statistics column.

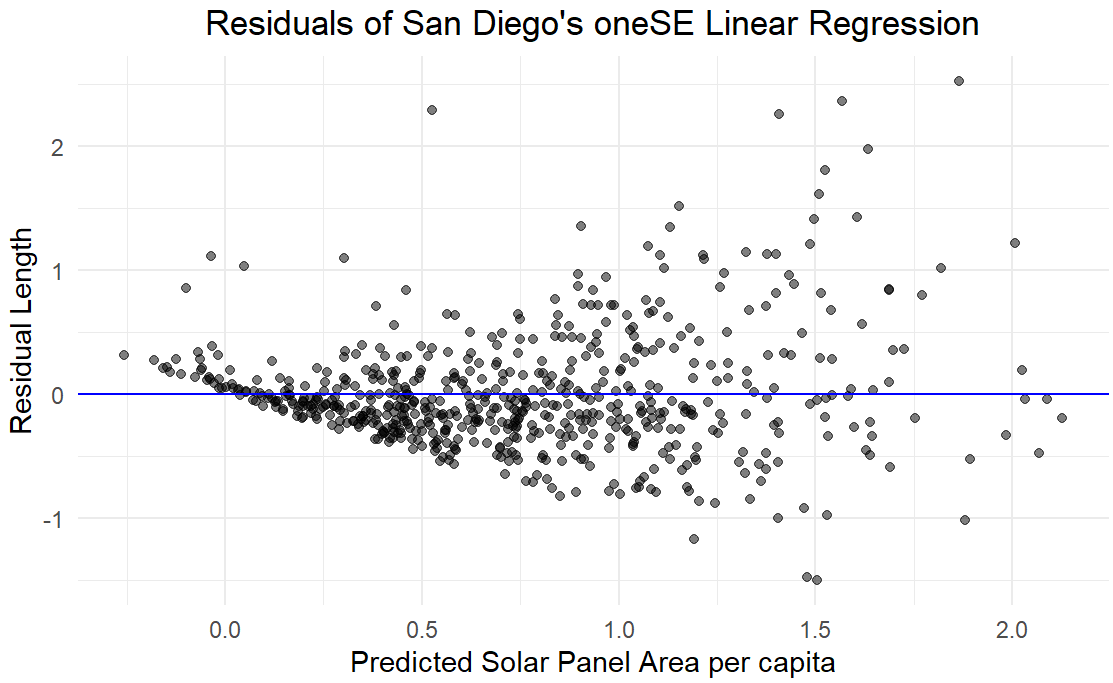

Figure 5: Residual length (calculated as the difference between actual solar panel area per capita and the predicted value based on the LASSO) plotted against the outcome variable (predicted solar panel area per capita) for $N = 619$ census tracts in San Diego County.

The LASSO results from Portland, Houston, and Chicago indicate less correlation than San Diego, as none of these three cities had a strong model to predict solar panel area per capita. Portland's LASSO selected only the intercept as a significant variable, meaning that there was so little correlation between the predictors and the outcome variable that simply using the average value of solar panel area per capita was within one standard deviation away from the best possible LASSO model. (Table 5) LASSO models for Houston and Chicago did include significant variables, even though their models were not strong. Houston's LASSO (Table 6) selected median age (+), GINI Index (+), and hypothetical metric tons of carbon offset (+) as the best predictors. Chicago's LASSO (Table 7) outputted the median housing unit value (+) and median yearly sunlight (+) as the best predictors. That said, we considered the R^2 values of these three models to be enough indication to halt further analysis of their residuals.

Portland LASSO Results			
Predictor	**Coefficient**	**Model Statistics**	
Intercept	0.108	**R^2**	NaN
		MAE	0.057

Table 5: Coefficients for the LASSO regression model for Multnomah County based on data inputted from *N* = 181 census tracts. No predictor variables are reported because a simple average of solar panel area per capita was within one standard deviation of the best possible LASSO model. R^2 (not applicable given the lack of a model) and mean absolute error of the model are reported in the model statistics column.

Houston LASSO Results			
Predictor	**Coefficient**	**Model Statistics**	
Intercept	-4.33×10^{-2}	**R^2**	0.09
Median Age	5.60×10^{-4}	**MAE**	0.01
Gini Index	7.83×10^{-2}		
Potential Carbon Offset (metric tons)	1.48×10^{-7}		

Table 6: Coefficients for the LASSO regression model for Multnomah County based on data inputted from *N* = 813 census tracts. Positive coefficients indicate that the predictor has a positive relationship with solar panel area per capita and negative coefficients show that the predictor has a negative relationship with solar panel area per capita. R^2 and mean absolute error of the model are reported in the model statistics column.

Cook County LASSO Results			
Predictor	**Coefficient**	**Model Statistics**	
Intercept	6.03×10^{-3}	**R^2**	0.07
Median Value of Housing Units	7.63×10^{-9}	**MAE**	0.009
Median Yearly Sunlight (kWh)	4.31×10^{-8}		

Table 7: Coefficients for the LASSO regression model for Multnomah County based on data inputted from $N = 1350$ census tracts. Positive coefficients indicate that the predictor has a positive relationship with solar panel area per capita and negative coefficients show that the predictor has a negative relationship with solar panel area per capita. R^2 and mean absolute error of the model are reported in the model statistics column.

Discussion

Both linear and LASSO regression modelling yielded significant results in only San Diego, where they show that race, median household income, annual sunlight level, median age, educational attainment, proportion of residents using a car as the sole mode of transportation, and the potential carbon offset are significant predictors of PV installation. Overall, the model for San Diego appears to be practical, as spatial factors play a role in determining PV installation rates. The significance of the social predictors in the regression models demonstrates that PV distribution in San Diego is not completely equitable, as census tracts with lower median household income, greater proportion of nonwhite residents, and lower educational attainment generally have lower levels of PV installation. However, there is still a large degree of unpredictability left unexplained by the model, as many census tracts have PV installation levels that differ significantly from the predicted values. This could be driven by factors exogenous to the model, such as commercial and military establishments in the county that may be equipped with many PV systems.

In each of the other three cities, both regression and LASSO modelling failed to show any significant relationship between the predictor variables, both social and spatial, and PV area per capita. We hypothesize that modelling primarily failed due to a lack of data to input into these models—Chicago and Houston have low PV installation rates (Table 1) and Portland has a comparatively lower population than other cities, meaning that there are significantly less census tracts than in the rest of the studied cities. This lack of data suggests that there might not be any significant trends in PV installation rates at its current level.

Overall, regression analysis of the cities studied shows that in San Diego, the city with high levels of PV installation and solar radiation, solar panel distribution is somewhat inequitable based on a variety of social factors. Cities with lower installation levels, such as Houston and Chicago, did not experience such trends. This suggests that broad expansions to solar panel installation may be inequitable, and cities that have not experienced such expansions have not

yet experienced such inequities. Thus, PV expansion in cities with low installation levels could learn from the example of San Diego. Future governmental programs could increase equity by encouraging PV installation in low-income communities of color and other traditionally underserved populations through economic incentives. Increased access would help these communities to potentially lower utility costs while helping cities to achieve ambitious carbon reduction goals.

Further research should investigate the claim that solar panel expansion has occurred inequitably in cities with high installation and solar radiation levels, as the inclusion of only San Diego is not enough to draw a conclusion about PV installation nationwide. Also, this research should consider urban geography more thoroughly, as military bases and large commercial establishments could skew regression modeling. Lastly, the presence of governmental incentive programs must be more carefully analyzed. These programs have the potential to vastly affect equity outcomes but were only considered in our dataset as a binary variable. The specificity and scope of these programs must be carefully considered to more fully understand the equitability of PV installation in the United States.

Works Cited

Barbose, Galen, et al. U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy, p. 11, *Residential Solar-Adopter Income and Demographic Trends: 2021 Update*.

Cook, Jeffrey J., and Lori A. Bird. "Unlocking Solar for Low- and Moderate-Income Residents: A Matrix of Financing Options by Resident, Provider, and Housing Type." *National Renewable Energy Laboratory*, 2018, https://doi.org/10.2172/1416133.

IRENA, "Renewable Power Generation Costs in 2020," 2021, https://www.irena.org/-/media/Files/IRENA/Agency/Publication/2021/Jun/IRENA_Power_Generation_Costs_2020.pdf.

Gernaat, David E.H.J., Harmen-Sytze de Boer, Louise C. Dammeier, and Detlef P. van Vuuren. "The Role of Residential Rooftop Photovoltaic in Long-Term Energy and Climate Scenarios." *Applied Energy*. 2020, https://doi.org/10.1016/j.apenergy.2020.115705.

Global Solar Atlas, Solar GIS, https://globalsolaratlas.info/map.

Google, Google, https://sunroof.withgoogle.com/.

Jakhongir Turakul Ugli, Tulakov. "The Importance of Alternative Solar Energy Sources and the Advantages and Disadvantages of Using Solar Panels in This Process." *American Journal of Software Engineering and Applications*, vol. 8, no. 1, 2019, p. 32., https://doi.org/10.11648/j.ajsea.20190801.14.

Matisoff, Daniel C., and Eric P. Johnson. "The comparative effectiveness of residential solar incentives." *Energy Policy*, 2017, https://doi.org/10.1016/j.enpol.2017.05.032.

Office of Energy Efficiency & Renewable Energy. “Residential and Commercial ITC Factsheets.” *Energy.gov*, 5 Feb. 2021, https://www.energy.gov/eere/solar/articles/residential-and-commercial-itc-factsheets.

O’Shaughnessy, Eric, et al. “Income-Targeted Marketing as a Supply-Side Barrier to Low-Income Solar Adoption.” *SSRN Electronic Journal*, 2021, https://doi.org/10.2139/ssrn.3826596.

“Solar Integration: Solar Energy and Storage Basics.” *Energy.gov*, US Department of Energy, https://www.energy.gov/eere/solar/solar-integration-solar-energy-and-storage-basics.

Stanford University, Stanford Engineering, Deep Solar Project. Retrieved February 21, 2021 from http://web.stanford.edu/group/deepsolar/home.html#

Sun, Eddie, et al. “Using Census Data to Predict Solar Panel Deployment.” Stanford University, 2018, http://cs229.stanford.edu/proj2018/report/104.pdf

United Nations. (n.d.). *The 17 goals | sustainable development*. United Nations. Retrieved February 21, 2022, from https://sdgs.un.org/goals

US Census Bureau. “2020 Census Demographic Data Map Viewer.” *Census.gov*, 8 Oct. 2021, https://www.census.gov/library/visualizations/2021/geo/demographicmapviewer.html.

US Census Bureau. “Urban Areas Facts.” *Census.gov*, 8 Oct. 2021, https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural/ua-facts.html.

US Department of Energy, Solar Energies Technology Office, Solar Futures Study (2021) Retrieved February 21, 2022, from https://www.energy.gov/eere/solar/solar-futures-study

US Department of Energy, Solar Energies Technology Office, Solar Soft Costs Basics (2021) Retrieved February 21, 2022, from https://www.energy.gov/eere/solar/solar-soft-costs-basics

Zsiborács, Henrik, Nóra Hegedűsné Baranyai, András Vincze, László Zentkó, Zoltán Birkner, Kinga Máté, and Gábor Pintér. “Intermittent Renewable Energy Sources: The Role of Energy Storage in the European Power System of 2040.” *Electronics* 8, no. 7 (2019): 729. https://doi.org/10.3390/electronics8070729.