1 Raw Data

The paper uses the following main data sets:

1. Deepsolar database: stored in Data/deep-solar
2. National Renewable Energy Laboratory's System Advisor Model (SAM): stored in Data/electricity-generation/sam
3. Google Project Sunroof: stored in Data/google-project-sunroof
4. Subsidy data: subsidy data from Sexton et al. (2021) are stored as Data/Solar_Incentives_Data_no_severin_data.xlsx
5. Open Grid Emissions version 0.3.0 hourly plant and power sector data for 2019 and 2020, along with the plant attributes tables which can be downloaded here. The data should be stored in Data/electricity-generation/open-grid-emissions
6. Tracking the Sun: December 2021 version, stored in Data/LBNL-pv-installations
7. AP3 Output: stored in Data/AP3

2 Code

There are 6 main steps to running the code. The steps should be executed in the order listed below.

1. Estimation of installation prices
2. Cleaning of electricity generation data and estimation of power plant production functions
3. Calculating hourly solar production profiles using SAM
4. Cleaning data for structural estimation and simulations
5. Structural Estimation and counterfactuals
6. Making final tables and graphs

2.1 Estimation of installation prices

We first use the Tracking the Sun data to estimate region level installation price functions. This is done in the R script R/install-price-panel-reg.R, which relies on the TTS data from December of 2021, saved as: "Data/LBNL-pv-installations/TTS_LBNL_public_file_14-Dec-2021_p1.csv" and "Data/LBNL-pv-installations/TTS_LBNL_public_file_14-Dec-2021_p2.csv", and the census region to state crosswalk in "Data/census-regions.csv".

The parameters of the installation price functions are stored as Data/CleanedData/state-install-price.csv.

The script also produces a Table displaying the results (tracking-the-sun-price-reg-2014-2018.tex), and a Figure (panels-price-region.jpeg) which compares the estimated fitted line against a flexible smoothing function for each region.

2.2 Power Plant Generation and Emissions

The code to clean the electricity generation data, estimate the power plant policy functions, estimate emissions rates and marginal damanges, and then produce model fit plots for the paper is all called from the R script R/run.R (note: it also is set up to run the installation price regression from the previous step and make the descriptive figures).

There are several data sources that are required for this :

- Open Grid Emissions version 0.3.0 hourly plant and power sector data for 2019 and 2020, along with the plant attributes tables which can be downloaded here.
- EPA eGRID data for 2019 and 2020, which can be downloaded from here
- EPA eGRID PM 2.5 estimated emissions rates from here
- EIA form 860 for 2019 and 2020, which can be downloaded from here
- EPA CEMS data for 2019 "Data/electricity-generation/cems/raw_cems_2019.fst"
- Table with stack heights from the EIA "Data/electricity-generation/plant-info/StackHeightInfo(20191202).csv"

The script first cleans the electricity generation data—collecting information about each plant (nameplate capacities, stack heights, locations), calculates hourly load and excess load in each NERC region from the OGE data, and then merges that with plant-level hourly production data. These collectively take 5 minutes.

Then we fit each power plant's generation policy function. This takes 25 minutes using 8 cores. The script then estimate marginal emissions rates and resulting marginal damages. PM 2.5 emissions rates are taken from eGRID. Marginal damages use the AP3 data. These take another 5 minutes.

Finally, there are a few scripts to visualize results. Two of them—the ones to simulate out-of-sample fit in 2020 and marginal damages across various levels of excess load take a considerable amount of time to run (200 minutes altogether), as we are generating hourly predictions from every power plant for many hours.

The outputs of this process—which are later used are all stored in "Data/electricity-generation/outputs". The power plant nameplate capacities are in "PPCap-oge.csv", generation models are saved as "PPRegressors-oge.csv", marginal damages are "damage-per-mwh-oge-scc-XXX.csv" (where XXX is the social cost of carbon in 2014 dollars), and hourly excess load by region in "RegLoad-oge.csv".

2.3 Hourly residential solar generation profiles

We use the System Advisor Model (SAM 2023.12.17 r1) to estimate hourly profiles of solar panel production by county, following Sexton et al. (2021).

The first step of this is to download weather data from Sengupta et al. (2018) for each county centroid. This is done using the NREL API from the script Data/electricity-generation/sam/01-nsrdb-weather-download-county.r, which saves the weather output in "Data/electricity-generation/sam/weather" and takes about 4 hours. Note, this requires NREL API key saved as environment variable "NREL_KEY".

You then open "simulate_generation.sam" in SAM, then open and run "GenerationScript_CEC_0-4999.lk" which simulates hourly generation for each of the county weather files, and saves the output as "intermediate_output/results_county.csv".

Finally, run "Data/electricity-generation/sam/02-make-county-profiles.r", which cleans the SAM output and saves it as, "Data/electricity-generation/output/solar-profile-county-dt.csv", and makes the figure with example hourly profiles in the Appendix.

2.4 Cleaning data for structural estimation and simulations

The Stata do-file StataDataCleaning/wrapper1.do takes raw data from Deepsolar, Google Project Sunroof, and cleaned data on electricity production and converts the data into .csv files to be loaded into Matlab. The file also runs some reduced-form regressions of installation and installation size on monetary incentives.

2.5 Structural Estimation and counterfactuals

The Matlab script StructuralCode/Main.m estimates the household installation component of the model, runs simulation, and calculates optimal policy within various environments.

To estimate the model, set the option (run_estimation=1) and run the code. The code will load the data, and estimate the structural parameters via maximum likelihood. The estimated structural parameters will be saved as Outputs/params_estCollPercPolDnm1.csv.

After estimating the model, to simulate the model given current subsidies, set (run_simulation=1) and run the code.

After estimating the model, to solve for optimal policy (run_opt_policy=1) and run the code. The options below control what type of optimal policy to calculate, and various counterfactual policy environments.

2.6 Making final tables and graphs

The Matlab script StructuralCode/make_tables3.m takes the output from the optimal policy calculations and creates the main results tables in the paper.

The Stata do-file StructuralCode/wrapper2.do uses the output from model simulations and makes the model fit graphs. This also creates the border-discontinuity graphs and runs border discontinuity regressions.

Finally, the R script R/result-maps.R creates the state-level result maps.

3 Computational Requirements

3.1 Software Requirements

1 Stata (last run with version 15.1)

2 Matlab (version 2022a on Linux)

3 R (version 4.3.1, 2023-06-16)

4 System Advisor Model (SAM 2023.12.17 r1)

All R packages are stored in an renv lockfile uploaded with this repository (renv.lock), to load all of the correct pacakge versions, run renv::restore().

3.2 Hardware and Runtime Requirements

The Stata code was run on a 4-core Intel-based desktop with Windows 10.0. The Matlab code was run on a 28-core Linux server with 128 GB of RAM and 200 GB of fast local storage. The R code was run on a Mac with an 8 core Apple M2 chip and 24 GB RAM.

The approximate compute times are given below:

- Estimation of installation prices $\approx$ 1 minute
- Cleaning of electricity generation data and rstimation of power plant production functions $\approx$ 4 hours
- Hourly solar profiles from SAM $\approx$ 4 hours
- Cleaning data for structural estimation and simulations $\approx$ 30 minutes
- Estimating Household Side of Model $\approx$ 1 hour
- Calculating optimal cost-neutral subsidies $\approx$ 20 hours
- Making final tables and graphs $\approx$ 5 minutes

References

Sengupta, Manajit, Yu Xie, Anthony Lopez, Aron Habte, Galen Maclaurin, and James Shelby. 2018. "The National Solar Radiation Data Base (NSRDB)." *Renewable and Sustainable Energy Reviews*, 89: 51–60.

Sexton, Steven, A Justin Kirkpatrick, Robert I Harris, and Nicholas Z Muller. 2021. "Heterogeneous solar capacity benefits, appropriability, and the costs of suboptimal siting." *Journal of the Association of Environmental and Resource Economists*, 8(6): 1209–1244.