Project Proposal

Gloria Li

Problem Statement

The energy grid is changing. For centuries we have seen the one-way flow of electrons, from generation to transmission to distribution, and the top-down concentration of power in the hands of the nation's electric utilities. With the rise of distributed generation (DG), an interesting new entity has emerged: the "prosumer" – a traditional consumer who now has the capacity to produce electricity (usually through rooftop photovoltaics) and sell their excess back to the grid. This exchange, called net metering, has generated considerable debate in recent years.

Solar advocates argue that distributed generation is environmentally friendly and provides ancillary benefits to the grid, while utilities argue that this causes inequitable cross-subsidization and incurs grid maintenance costs. In a 2017 study, 59% of utility executives agreed that small-scale DG places stress on their network hosting capacity and reduces revenue, forcing them to raise rates on other customers and turning more customers toward DG[1] – the so-called "utility death spiral".

The high-level research question I hope to explore with my analysis is: is more DG (i.e. net metered capacity) correlated with changes in grid reliability over time in the United States?

Data

The data source I intend to use is the Energy Information Administration's Form 861, also known as the Annual Electric Power Industry Report. It comprises data collected from approximately 3,300 U.S. respondents, including electric utilities, wholesale power marketers and electric power producers.

The primary variables I will analyze from EIA-861 include net metering data from 2007-2019 as the predictor and grid reliability data from 2013-2019 as the response (years reflect all available data). Other variables available from this source that I may decide to incorporate in my analysis include demand response programs (2013-2019) and energy efficiency (2013-2019). These are all potential sources of endogeneity that may help to provide a more accurate view of the relationship between net-metered distributed generation and grid reliability over time.

The datasets for each year are zipped and publicly available on the E.I.A. website.[2] I will be downloading them from there, but there is a considerable amount of data wrangling I will do to produce a final dataset I can work with to answer my question (see description in next section). I plan on incorporating at least a few, if not all, of the following variables (divided by the dataset where they can be found):

- Reliability
  - System Average Interruption Duration Index (SAIDI), with and without Major Event Days: I will look at both because SAIDI without MED excludes low-frequency, high-impact events such as hurricanes.[3]
  - System Average Interruption Frequency Index (SAIFI), with and without Major Event Days
- Net metering – All Technologies (Photovoltaic, Wind, Other)
  - Total installed capacity for the Residential, Commercial, Industrial and Transportation sectors (MW)
  - Energy sold back to grid (MWh)
  - Storage Capacity (MW)
- Demand Response

  - Actual Peak Demand Savings (MW)
- Energy Efficiency
  - Incremental Life Cycle Energy Savings (MWh)
- Peak Demand Savings (MW)

Methodology

This data is zipped by year, so I will need to clean it to make it tidy and then merge all the years together into a large dataset, most likely joining by electric utility. This will include tasks such as: reformatting the data for reading into Python, assigning utility identifiers, changing variables to a per capita basis if necessary, and doing energy capacity conversions (MW/MWh). If I use the data from 2013-2019, this new dataset will join together 4 datasets (described above) for each year.

I am planning to make multiple data visualizations using Matplotlib to present the relationship between the increase in net metered capacity and the changes in grid reliability over time.[4] These include a multi-line graph that shows the grid reliability and the net metering capacity over time and a scatterplot that includes a fitted regression line.

We have not gone over how to manipulate spatial data in class, but if I have time, I would also like to make a map of the United States that shows the 2019 (a.k.a. most recent) grid reliability and net metered generation. One note of caution here is that the utilities are not neatly divided into states; there are thousands of utilities and service territories that are piecemeal across the states. One way to tackle visualizing them may be aggregating the service territories by state.

For my machine learning component, I will focus on using OLS to find the relationship between net metered distributed generation and grid reliability. I will use the Scikit-Learn library and the cross-validation techniques we learn in class to split my data up into a test, training, and validation set.[5]

Evaluation

Success for my project will primarily that I have run a robust OLS analysis on my reformatted data that can be replicated and defended. On the machine learning side, if the approach I take reduces the MSE past an ordinary OLS, that will be a good learning opportunity too. I know this is an ambitious project to tackle, and it is important to me personally to give adequate treatment to other relevant variables in order to reduce endogeneity, especially as I recognize the policy and political implications of certain conclusions. I intend for the visualizations to augment the story that the OLS tells and make it more digestible/convincing, especially to a hypothetical audience that may consist of anyone from a solar advocate to a utility executive.

I will know that I accomplished my goal if I have found evidence of a meaningful correlation, either positive or negative, that can shed light on how to improve grid reliability moving forward. It will be equally successful if I find evidence that there is no correlation, as long as that makes sense given the data – although in this case, I will need to think more about how to interpret that result.

Works Cited

1. John, J. (2017, June 06). Utility Execs See Distributed Energy as the Biggest Stress on Grid Reliability, Revenues. Retrieved October 30, 2020, from https://www.greentechmedia.com/articles/read/utility-execs-see-distributed-energy-as-biggest-stress-on-grid-reliability
2. U.S. Energy Information Administration - EIA - Independent Statistics and Analysis. (2020, October 6). Retrieved October 30, 2020, from https://www.eia.gov/electricity/data/eia861/
3. Susser, J. (2018, July 25). Understanding and Managing Grid Reliability and Resiliency. Retrieved October 30, 2020, from https://www.advancedenergy.org/2018/07/25/gridreliabilityandresiliency/
4. John D. Hunter. Matplotlib: A 2D Graphics Environment, Computing in Science & Engineering, 9, 90-95 (2007), DOI:10.1109/MCSE.2007.55

5. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay. Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825-2830 (2011)