Jacob Gratrix

Final Project Proposal

Stat 4185 – Sec 001

First I wish to outline the phenomena at hand for this project. Many don't realize that the US power grid is the largest single piece of infrastructure on earth. I'm curious about the distribution of electricity markets in the US with respect to net generation and net consumption. This is an intelligible question because the US is split into regions governed by ISOs or independent system operators, who buy and sell electricity on the interstate wholesale market. They in turn regulate a lot of the bidding activity of individual electricity production firms and power grid mantainance firms (whom consumers interact with). This means that there is publically available data about the price of electricity at any given time. Furthermore, there is accessible data on power generation methods by area, be it nuclear power, fossil fuel power, or renewable resources like hydropower. And these are obviously not evenly distributed from region to region.

Allow me to further explain my logic with some cases. In regions with a greater rate of nuclear power generation for instance, it would be reasonable to surmise that the price of electricity *from* that region correlates positively with the relative demand in neighboring regions. Meanwhile this sort of analysis tranfers to the consumption of electricity. For instance in regions with a higher than average level of urbanization *and* commercial/industrial activity, we could reasonably expect the cost of adopting renewable power generation to be greater. This is because industrial firms tend to be low-variance consumers of large amounts of electricity, while renewable sources of electricity generation are high-variance producers of lower amounts of electricity. So this particular incentive structure is negatively favored, and therefore results in a system further from an optimal state of electricity supply matching demand exactly. One might wonder how large of a deadweight loss this is on the economy of different regions. If my memory serves, in recent months there was grief on the West Coast and in the American Southwest about rolling black-outs. These were in part caused by one of the mismatches detailed above.

Now I want to articulate my proposed questions. If I am able to reasonably estimate how optimally the current grid satisfies the needs of the populace, then I will consider this project extra fruitful. A first question is *which regions produce more than they consume and which consume more than they produce?* This is a necessary step since any reasonable analysis of the entire US energy market has to somehow discretize the search space over the decision variables. In this case, this means I need to analyse regions rather than singular nodes and properties/households.

My second question is as follows. *How does a region's classification as a 'net producer' or 'net consumer' relate to population/urbanization, geography, prevailing climate, grid connectivity, and energy source allocation?* I might even get a bit ambitious and try to add public policy variations in my analysis, for instance taxes and subsidies on the consumers and producers of electricity. Of course this could backfire due to the complexity of investigating this variable and attempting to quantify it in a uniform manner.

As far as methods are concerned, I plan on using a couple data gathering methods. There are excellent APIs for retrieving electricity price data for ISOs, and that's probably the only viable way to get the data I need. I can implement web scraping to ascertain the power generation methods of every region, and some other variables like population and key metrics related to climate. Commercial activities by region might also be attainable using web scrapped data. The 'grid connectivity' variable I mention is not that easy to grasp. I'm referring principally to the *ease* of transacting electricity between neighboring regions, or the *difficulty* of transacting with non-contiguous regions. That is something I will need to research further.

The precise models I end up using will depend on (a) the formats of my variables, quantitative versus qualitative, continuous versus ordinal/discrete, and (b) how nonlinearities spring up in the relations among

the variables. For instance I wouldn't want to apply an naiive multiple regression to data which are evidently non-linearisable, such as exhibiting extreme heteroskedasticity. Nor would I want to employ any ensemble models without a way to somehow 'weight' the subjects e.g. regions with differing populations or differing climate profiles. I'm less worried about the ability to create visualizations for my data, as the subject matter at hand will lend itself to some interesting visualizations regardless of my ability to model the relationships involved.

I welcome any comments, feedback, or tips regarding these ideas for my project. As a huge infrastructure nerd, I think it would be super cool to get insights from these proposed data and methods.