Notes from July 21st:

I tried to figure out how to calculate customers using the EIA data today. I used Vivian and Nina’s notes, and I read a lot of stuff. Here are two resources that I drew on extensively:

Federal energy regulation commission’s energy primer and other infographic:

<https://www.ferc.gov/sites/default/files/2020-06/energy-primer-2020_0.pdf>

<https://www.ferc.gov/electric-power-markets>

Idk what this is but it was super helpful:

<https://www.rff.org/publications/explainers/electricity-101/>

<https://www.rff.org/publications/explainers/us-electricity-markets-101/>

Some random person’s writing about working for/being in management of a power marketer:

<https://nap.nationalacademies.org/read/5482/chapter/8>

I read about this with the goal to figure out what to include and exclude in total state customer counts derived from the detailed data spreadsheets available on the EIA website. EIA also provides straight-up state-level counts of customers that are available for download, and I think that these would be reasonable to use for estimating state customers, after reading about what they are. However, we also need to understand the detailed data because we want to get customer estimates only for the utilities included in POUS, in addition to those total state counts.

Recall: we want to estimate the total number of customers in a state, because we want to create estimates of how many customers POUS covers by county. We’ll do this by scaling down the state-level estimates of electrical customers, according to the proportion of households and establishments estimated by the census in each county, and then we’ll compare that number to the number of POUS customers in the state. We’ll get the number of POUS customers in the state by linking utilities included in POUS with EIA, and excluding those not included. Hopefully we can start with the POUS utility names and match most, if not all, of them.

So the data from EIA: EIA collects reporting about sales, customers, etc. from all companies involved in energy markets. This doesn’t mean just vertically integrated utilities, or transmission and distribution utilities - they also collect data from intermediaries like retail power marketers. It seems at first glance, from the dataset and from the form that utilities have to fill to actually give EIA the data, that the customer counts by utility may include double counts, since companies involved in different parts of the market (EX wholesale vs distribution) may cover the same customers.

The “Part” code in the dataset and the ‘Ownership’ description in the dataset both give us information about what a company does in the energy market.

We know that there are vertically integrated utilities, which generate and distribute electricity, and this is what EIA defines as “bundled service” – generation and distribution. These are typically part of regulated markets, but not always, and coops are usually examples of these. These would fall under part A. There are companies that only generate energy and sell to other companies, in deregulated markets, which would be part B. This category also includes behind-the-meter entities. These are things like solar farms and rooftop solar projects, which offset energy costs for their owners by sending electricity back on to the grid, but don’t like, sell electricity to customers. IDK what their customers counts mean. So, they shouldn’t be included in total customer counts. Part C includes distribution-only utilities, again usually in deregulated markets, and then Part D includes bundled service from companies that seem to be both retail power marketers and energy service providers. This is only for Texas. Maybe this is so because this combination of retail power marketing and also distribution is a think unique to Texas: these companies don’t do generation (that would be part A), but purchase generated electricity from other companies and also own transmission and distribution stuff.

Also, I think that companies can fit into more than one of these categories, such as Nueces Co-op. I think it’s possible this coop generates, and also purchases electricity wholesale, and distributes, putting it in parts A and D, and reports how many customers it serves under each model. I think including both of these numbers in the customer totals makes sense.

EIA tells us in the datasets that in order to avoid double-counting, we should not include part C, which is distribution only. I think this saves us from counting customers that are covered by generating utilities and distribution utilities. It also says it excludes behind-the-meter reports, which are the rooftop solar customers, from state totals, but tells us to calculate simply by summing what’s in the spreadsheet. I think this is because the data includes and adjustment, which is for non-response of certain utitlies and ‘customer-sited utilities’, which means rooftop solar. So all in all, we just need to do what it says at the bottom of the spreadsheet to do.

I will calculate totals using Parts A, B, and D.

More notes July 12th:

I’m going to write a bit about each of the other scripts, and the data cleaning so far.

The data have come to us by sea…JK this is not an Isabelle Allende novel. The data have come to us in city-county-utility units. The ‘city’ I use here is loose – these city names are really sometimes neighbourhoods, towns, half of a city, etc. They are sub-county areas. The observations in the dataset are supposed to be time series within these areas.

However, the data are not stored in a time series format. Within each city-county-utility area, there are observations recorded only when there is a change in the number of customers out, or when there is missing data, and these records are time stamped. Missing data is represented by two matching time stamps. Need to explain this more.

02\_expand.R expands the outage data to a timeseries. This takes a lot of time and computing power (and storage too) but I think it’s important because we need to identify how complete our data is, in order to make good, rigorous decisions about its use.

So, this script expands the data to a timeseries at 10-minute intervals, starting at the first observation present for a city-county-utility unit, and ending at the last observation. This is a choice, because we could expand to the time period that we think the data should cover and that POUS says it should cover, but there are a lot of places that have very few observations. In fact, the distribution of observations by city-We don’t want to make the mistake of interpolating coverage in these places.

We take into account the missing observation markers in the data by carrying forward observations to missing data markers, and then leaving data missing after such markers. There are cases where, if there are only a few observations in one area, if we do this, we’ll carry forward ‘complete data’ to the end of the study period because *the missing data markers are not there*. I’ve avoided this problem by only doing this operation in between observations.

03\_hourly\_locf.R:

We carry forward missing data only if the period between two observations that are non-missing is 4 hours or less.

We want to change this because the idea here is that, we want to carry forward when we think interpolation makes sense, and we want to not when it doesn’t make sense. We can simulate this threshold.

Notes from Heather’s data work on Jun. 21st:

Today I finished cleaning the dataset so that we have clear county FIPS information. I did this iteratively, because there are many reasons why the county information in the dataset may not readily admit a FIPS code. I tried to do the easiest things first, and then move to more labour-intensive stuff.

Right now this is the first script in the data processing pipeline.

I’m focusing on assigning FIPS based on an NHGIS dataset both because a) I don’t trust POUS to assign FIPS correctly, b) a lot of places are missing FIPS codes anyway, and c) we need to map these things at some point and know what these locations are, so we may as well match them to the NHGIS FIPS codes now. I also want to have the dataset be so that the county and state names really do correspond to a location and using them won’t mess things up. This makes it intuitive.

First, I cleaned state and county names for punctuation, spaces, capitalization, etc. and removed words like ‘borough’, ‘township’, ‘county’, etc. I checked that this cleaning did not make state-county name combinations less unique than they were before.

I matched these state-county name combinations with a NHGIS shapefile containing FIPS codes. This was the first pass.

I also filtered out observations where the county name is a number which doesn’t make sense (I checked to make sure these were not FIPS codes), or a word that doesn’t make sense (some county names are the word ‘Zip Code’…..? Nice data error, POUS.) We should contact POUS about the state of Kentucky, since we have almost no observations there, as all the county names are non-FIPS numbers. It would be nice to match these so that we can cover Kentucky.

Then I did some trickier stuff.

I created a file which matched names spelled differently in the POUS dataset with the NHGIS file so a FIPS could be assigned.

Some observations still missing didn’t match because the state-county combination didn’t exist in reality and therefore didn’t’ exist in the NHGIS data. We think this happens when a utility primarily serves one state, but also serves some communities just over state lines. We think this is often recorded “STATE WHERE WE NORMALLY SERVE”, “ACTUAL COUNTY WE’RE IN, IN A NEIGHBOURING STATE”. I created a file where, if a state-county name combination didn’t exist in reality and the county name listed was in fact a county in a neighbouring state (and only one neighbouring state), I reassigned the state information to match along with a correct FIPS code.

There are still some state-county combinations that don’t exist, which could probably be matched to the correct counties if we looked up the utility companies and the areas they serve. These make up the majority of the missing data remaining that we could match.

However, given what I’ve done so far, we’ve matched 98% of observations and 87% of initial state-county combinations, some of which were state-county combinations where the state or county was listed as unknown or blank. We’ve also matched 97% of counties in the US. I think this is good enough! (Except for Kentucky.)

I also did check to see if any of my assignments disagreed with POUS. They did not, so that’s great, (and may say something about the correctness of the above state-reassignment strategy), and I just really added additional ones. I should check if I dropped any that did have FIPS codes. Haven’t done that.

It is possible to match some of these utilities with Unknown county info to a county by searching the utility online and finding out, but I thought that was too hard.

Notes on the Power Outages US (POUS) dataset, and plans for simulation and analyses from our (Marianthi, Joan, Dan, Heather, Francesca was absent due to COVID ☹) meeting on May 2nd.

The original goal of the meeting was to discuss simulations. There are several sources of uncertainty in the data which we want to explore with simulations.

Preamble about the data structure: POUS gives us information on the number of customers out every ten minutes by utility and ‘subcounty area’. These subcounty areas can be cities, neighbourhoods, or towns, etc., and there are examples of cities, towns, and neighbourhoods in the dataset. They are identified in the dataset by the county name and subcounty area name (for example, “Los Angeles County, Los Angeles Department of Water and Power, West LA”). We call these ‘power operating divisions’ or PODs. These are the smallest geographic units we have.

A ‘customer’ is one electrical meter. One customer can therefore be a household, several residential units in a building all metered together, one residential unit in a building metered separately, or a business like a restaurant or industrial complex. POUS does not reliably provide the number of customers (or people) served in each POD, though it does include (bad) estimates. We also do not know the exact locations of each of the PODs, because POUS doesn’t provide anything like a shapefile defining POD boundaries – we only know the county, utility, and subcounty names they give us. We have estimated the number of electrical customers by county using census data on household and establishment counts.

Because we can’t locate PODs exactly and only have denominator estimates at the county level, we think conducting analyses at the county level is a good idea.

What we want to simulate:

1. Missing data:
   1. The POUS dataset relies on utility company website APIs to retrieve information about the number of customers out in a POD. These APIs sometimes don’t respond for minutes, days, or weeks, etc., meaning we have periods of time with no data. Some PODs have almost no data – they have only 1 or 2 10-minute periods where the API *is* responding, and the rest of the time indicate non-response. We want to know when to exclude PODs from the analysis because they don’t have enough data. We may pick a threshold after which, if more than a certain percentage of the study period is missing, we exclude a POD from out analysis. We want to find a threshold where we still cover most of the US, but where the remaining data, though it contains missingness, is still reasonably reliable (aka, our analyses are still able to detect associations between exposure and outcome if they are present). We also want to assess how using LOCF could influence our results.
   2. Not all utility companies have APIs or websites, and so some are not included in the dataset. This means not all people in the US are covered by the dataset. Additionally, it’s possible (and frequently true) that more than one utility company serves the same area, and some utilities serving a certain area may be included while others are not. Because of that, and because we can’t locate these PODs precisely, finding which geographic areas are missing from the data is tricky. We may need to exclude counties where more than a certain percentage people are *not* tracked by POUS, or somehow extrapolate data from those who are tracked, to the whole county. We want to use simulations to decide on a coverage threshold after which to exclude counties. Separately, we should assess the correlation between outages in utilities that serve the same county, which would help us decide if extrapolating information from customers who are tracked to the whole county is a reasonable way to deal with missingness. This is not a simulation question, just a thing we need to do separately.
   3. Note: This last point is particularly sensitive because rural utilities are less likely to be included in the dataset. Rural residence may be associated with hospitalization for CVD, an outcome we are interested in. If we overestimate or underestimate rural exposure this could produce confounding.
2. Exposure definition:
   1. We also want to use simulations to assess how best to define our exposure. Previous papers have created a binary exposure, saying that a power outage happens when a certain percentage of customers served in a geographic unit are out. We want to experiment with the threshold – what percentage of customers out makes the best threshold for a power outage? This can also help us interpret previous studies.
   2. We also want to try out a continuous exposure where we measure the percentage of customers out without changing it. This may be harder to communicate (we can’t say that RR of hospitalization increases x amount with a power outage if our exposure is continuous – the conclusions of the study would be harder to communicate)

Other notes:

We plan to use a big DiD for our analysis similar to Francesca, Joel, and Dan’s paper. This relies on the assumption that power outages are rare and are unlikely to happen again at the same time between years. We should check this assumption for whatever definition of power outage we decide to use.

We should also assess if SES is correlated with household size. Because we have data at the household level, if households are bigger in lower SES areas, we may accidentally underestimate exposure in these areas.

Notes from my discussion with Marianthi about the simulation plan:

(This was just our very first thoughts on what might be good to start with).

For our first pass, let’s look at the association between power outages and CVD.

Also for first pass, want to change how we define a power outage (by percent of customers out – 10% is PO, 20% is PO, etc.)

Need to find a reasonable RR for association between PO and CVD based on the literature, for example RR = 1.2 per one additional power outage (daily exposure)

Let’s say we simulate 1 year of data (we’ll be looking at 1 year of data for medicare paper)

1. simulate 100 PODs & for each one we simulate number of customers based on the observed distribution of customers we see in our data (need to find this)

2. Need to make an assumption on frequency and duration of actual POs — this will be from data, but will also vary based on our PO definition. Perhaps should base this on the distribution of customers out in a POD instead??

3. create different datasets with different PO exposures based on different cutoffs of customers experiencing a PO — e.g., 10%, 20%, etc — and the assumed distribution of customers out (from previous step)

4. Need to check the distribution of customers vs individuals (from census) etc — so our different datasets have both columns (customers & individuals) — that would help us get an idea of exposed individuals (based on cutoff) — this should be a function of steps 1 & 3

—> Now we should have exposed individuals per day per POD (a) if at least cutoff exposed, assign everyone exposed vs (b) only cutoff exposed

So then we should have

POD ID, Date, Individual ID, assumed total exposure (#a from 4), true exposure (#b from 4)

Assign outcome based on assumed RR and true exposure + random error on each individual

For surrogate and true exposures run a logistic regression by subject [with random intercepts for POD ID and individual ID?] — or set it up as a case crossover? (conditional logistic)