eGRID Electrical Generation and Emissions Data From U.S. EPA And ACS Small-Area Socioeconomic Data from the U.S. Bureau of the Census: Tools for Environmental Justice and other Socio-Technical Energy Research in the United States

Michael Ash*

September 11, 2023

Learning objectives, directions, and scavenger hunt for session. Introduction to Electrical Generation and Emissions Data from U.S. EPA. Introduction to small-area data from the U.S. Bureau of the Census. The code and data for this introduction are available from the GitHub repository `https://github.com/michaelaoash/ELEVATE-census-egrid-intro`.

- **Learning Objectives and Skills**
- Learn about the U.S. EPA Emissions & Generation Resource Integrated Database (eGRID) including the metadata (data sources, organization, observations, and variables).

*Department of Economics, School of Public Policy, Political Economy Research Institute, University of Massachusetts Amherst, `mash@umass.edu`

- Learn about the Air Pollution Emission Experiments and Policy (APEEP) model to monetize damage estimates from eGRID emissions
- Download eGRID and load eGRID into R, load APEEP into R, join eGRID and APEEP data
- Filter and tabulate eGRID data
- Learn about the American Community Survey of the U.S. Bureau of the Census including the metadata (data source, organization, observations, and variables) and related geographic data (TIGER/Line Shapefiles)
- Use R packages `tidycensus` and `tigris` to download, filter, join, summarize, and map and label small-area socioeconomic data and related geographic data for Holyoke, Massachusetts.
- Use GIS functions in R to overlay Massachusetts electrical-generation facilities with Holyoke geography and compute the distance between each electrical generation unit and the wards of Holyoke.
- **Questions for Reflection**
- What social, economic, environmental, or energy variables might be useful on a small-area basis? What are questions that we might answer with small-area data? What kind of relationships can we explore with small-area data, for example, between toxic exposures and test scores, between poverty rates and energy bills, between voter participation and enforcement of environmental regulations?
- Are some of these variables available from eGRID, ACS, or other datasets? How do you find the documentation?
- What are some of the advantages and disadvantages of the Census small-area geography of Tracts and Block Groups? What are some alternative geographies for small-area analysis, and what are the advantages and disadvantages of each?
- In our maps, the entirety of each Block Group ("neighborhood") was coded with one value of the poverty rate. Why might that pose a problem for some applications?

- eGRID lists electrical generators. How complete is the coverage? (Can you name a generator that is not included? How do tabulated totals, e.g., for generation, compare to reported totals from other sources?) The eGRID data are annual; is that frequency appropriate for the questions that you investigate?

1. Advance preparation. These exercises require an installation of the R statistical software (`https://www.r-project.org/`), including the `tidyverse` suite of packages, the spatial and GIS packages `sp` and `sf`, and the specialized Census data packages `tidycensus` and `tigris`. Please install R and these packages. (R and all of the packages are free, open source, and available across platforms. ELEVATE peers or mentors may be able to help with installation.)

2. Visit the website for the "historical" eGRID data `https://www.epa.gov/egrid/historical-egrid-data`. We will use the 2020 data because supplemental data on Particulate Matter emissions (`https://www.epa.gov/egrid/egrid-related-materials`) are available for 2020 (but not yet for 2021). Download the metric and non-metric Excel data. Also download the eGRID PM2.5 Data. (The downloads are already included in the GitHub repository, and you can skip the download step if you have already downloaded the repository.) Open the spreadsheets in Excel or another spreadsheet program to examine the data, but *do not modify* the data in any way.
   (a) eGRID integrates data on electrical-generation facilities US DOE Energy Information Administration and the US EPA Clean Air Markets Division. What is the difference between the Unit, Generator, and Plant sheets? (We will focus on the Plant sheet.)
   (b) What variables are available on each plant? Review some of the columns on the Plant sheet of the eGRID workbook. Find several columns that concern location and management. Find several columns that concern electricity generation. Find several columns that concern pollutant emissions.

3. Make sure that the script files and the egrid spreadsheets are in the same directory. In the same directory are several files from the Air Pollution Emission Experiments and Policy analysis (APEEP) model developed by Nick Muller at Carnegie Mellon and others `https://public.tepper.cmu.edu/nmuller/APModel.aspx`. These files compute monetized estimates ($/ton) of the damage to human health, agriculture, and physical capital per ton of emissions based on the county of *release*.

4. Open the script file `egrid2020_metric.R` or `egrid2020_nonmetric.R` (your choice) in Rstudio or another integrated development environment (IDE) for R.

Work through the script one line at a time, carefully reviewing inputs and outputs. Find, run and review the output of the summarize routine that computes shares of capacity, generation, and emissions by fuel.

5. Find the routine that looks for generation that may be associated with universities. Modify the search terms to see if you can find additional university-based facilities.

6. At the end of the program we focus on Massachusetts facilities and save a dataset of fossil-fuel facilities in Massachusetts and Connecticut.

7. In the `tidycensus-demo/holyoke-poverty` directory of the GitHub repository, open the R script file `holyoke-poverty.R`. This file demonstrates the use of the `tidycensus` package in R. Please go to the link and get your own Census API key and replace mine with yours.

8. We will use the American Community Survey (ACS) 5-year data (pooling survey results from 2016–2020, which permits analysis of small areas such as Census Tracts and Census Block Groups. The ACS is a household survey that asks each sampled household questions about demographics and income. We will focus on the determination of poverty based on household composition and income and compute poverty rates for families with children for Census Block Groups (similar to neighborhoods) in Holyoke.

9. First we find the relevant variables. We download and store the codebook for the ACS with the `load_variables()` function in `tidycensus`. The codebook has four columns: the **name** of each variable (typically a letter followed by 5 numbers, sometimes an additional letter, followed by an underscore followed by 3 more numbers); a **label** that explains exactly what the variable expresses; a **concept** that describes the broad area of interest, and **geography** that specifies the smallest area for which the variable is available. We can filter the codebook with the `filter()` function and use `grepl()` to identify words or phrases of interest. The syntax is `grepl("text to search for", variable to search for the text)`. There are roughly 30,000 variables in the ACS; most of them are *counts* of people in a specific category but some are average or median values. Record the names of the variables that you need. **Challenges:** How many variables are available only down to the Tract level, and how many variables are available down to the Block Group level? What variables related to household energy use are available from ACS? (What search terms did you use?)

10. Then we use `tidycensus` to download the needed variables directly from the ACS website at the U.S. Bureau of the Census, and we compute the poverty rate

from the poverty counts. It is always helpful to look at the data at intermediate steps. Geographic data which enables map plotting is included with the economic and demographic variables. **Challenge:** Are data available for Puerto Rico?

11. How big are Block Groups? Look at the map of Holyoke with Block Groups indicated. If you have been to Holyoke, estimate how long would it take you to walk across a downtown Block Group. Is there variation in the area and population of Block Groups? How many households are there in some of the Block Groups of Holyoke?

12. Look at the coding of the Census geography. The GEOID for a Block Group is a 15 character string: `SSCCCTTTTTTB`. This looks daunting but it can be easily parsed to show the geography:

`SS`	Two-digit code for the state in more or less alphabetical order. Massachusetts is 25.
`CCC`	Three-digit code for the county within the state (usually only odd numbers). Hampden County is 013.
`TTTT.TT`	Six-digit code to identify the tract within the county. The . is not shown, but the first four digits are usually sufficient to identify the tract and the last two digits are often 00. Sometimes the last two digits, e.g., .01 and .02, are coded if a Tract splits over time. Tracts are selected to have between 1,200 and 8,000 people with a target population of 4,000.
`B`	One-digit code to identify the block group with the tract. There are typically 1–4 block groups in a tract.

    While counties are a political jurisdiction (although not much of one in Massachusetts), Census Tracts and Block Groups are not jurisdictions and they may even cross the lines of municipalities. Tracts and Block Groups are typically selected to conform to physical boundaries (rivers, large roads, etc.) and social boundaries (roughly corresponding to "neighborhoods"), but there is variation over time and space. Block Groups partition Tracts which partition Counties which partition States, but the connection to other relevant geographies (wards and precincts, Congressional and other legislative districts, School districts, Zip codes, municipalities, etc.) is not always crisp. These other geographies are often available (from the Census or other public and private sources), and they can sliced up to match Census geography, but the process is more complicated.

13. We also use several TIGER/Line files to decorate the map with water, some large streets and roads, and voting wards. These can be separately downloaded from

the Census or accessed using the `tigris` package.

14. The script culminates with a poverty-rate **choropleth** of the Block Groups of Holyoke.

15. A coda returns to the eGRID data, converts the latitudes and longitudes in the Massachusetts fossil fuel electrical generation data, plots a map of the generators within the cities and towns of Massachusetts, and then computes the distance between each generation facility and each ward (voting district) of Holyoke.

16. Find a fossil-fuel EGU that is within Holyoke. What ward is it in? What was its primary fuel, and how much energy did generate in 2020? Find the next closest fossil-fuel EGU and report on it.