Energy Data Documentation

1 Data Sources

1.1 Employment and earnings

- **Source:** Decennial Census (1980 5% sample, 1990 5% sample, 2000 5% sample); American Community Survey (2010 3-year file, 2017 5-year file, 2022 5-year file)
- **Download link:** All data are downloaded from IPUMS (https://usa.ipums.org/usa-action/variables/group)
    - 1980 5% sample downloaded on May 9, 2022
    - 1990 and 2000 5% samples downloaded on April 24, 2022
    - 2010 3-year file downloaded on July 10, 2024
    - 2017 5-year file downloaded on September 24, 2024
    - 2022 5-year file downloaded on July 30, 2024
- **Variables used (raw variable names from IPUMS download):**
    - **gq:** Group quarters status
    - **age:** Age
    - **educ:** Educational Attainment
    - **perwt:** Person weight
    - **empstat:** Employment status
    - **wkswork2:** Weeks worked last year
    - **uhrswork:** Usual hours worked per week
    - **incwage:** Wage and salary income
    - **incbus00:** Business and farm income
    - **race:** Race
    - **hispan:** Hispanic origin
    - **bpl:** Birthplace
    - **citizen:** Citizenship status
    - **sex:** Sex
    - **statefip:** State FIPS code
    - **puma:** Public Use Microdata Area
    - **occ:** Occupation code
    - **ind1990:** Industry Code

1.2 Renewables potential

- **Source:** State and Local Planning for Energy (SLOPE) platform by the National Renewable Energy Laboratory (funded by the DOE).
- **Download link:** Data are downloaded from the platform's data viewer (https://maps.nrel.gov/slope/data-viewer)
    - Downloaded on October 29, 2024

- **Variables used (raw variable names from download):**
    - **geographyid**: County FIPS code
    - **countyname**: County name
    - **technology**: Type of renewables (land-based wind, utility photovoltaics)
    - **technicalgenerationpotentialmwhm**: Technical generation potential (MWh)

1.3 Electricity Generation

- **Source:** U.S. Energy Information Administration Form EIA-860
- **Download Link:** https://www.eia.gov/electricity/data/eia860/
    - Downloaded between October 11 and October 30, 2023.
- **Variables used (raw variable names from download):**
    - **Year**: The year of the data observation.
    - **Status**: Operational status of the generator (e.g., Operating, Standby).
    - **Plant ID**: Unique identifier for the power plant.
    - **Generator ID**: Unique identifier for the generator within the plant.
    - **Energy 1**: Primary fuel or energy source used by the generator (e.g., Fossil Fuel)
    - **Energy 1 Sub-Category**: More detailed classification of the primary fuel or energy source (e.g., Coal, Nuclear, Biomass).
    - **State FIPS**: FIPS code for the state where the generator is located.
    - **State**: Abbreviation of the state where the generator is located.
    - **State Name**: Full name of the state where the generator is located.
    - **County FIPS**: FIPS code for the county where the generator is located.
    - **County**: Name of the county where the generator is located.
    - **Nameplate Capacity**: Maximum output capacity of the generator under specific conditions (measured in megawatts).

2 Data Preparation

Key datasets

- **Commuting Zone Crosswalk, Industry Crosswalk, Occupational Code Crosswalk:** Crosswalks with 1990 values published by Professor David Dorn (link).

2.1 Employment and Earnings

- Because county information isn't consistently available across ACS/Census files, we crosswalk on Public Use Microdata Areas ("puma"). Using the state FIPS code and the "puma", we crosswalk to commuting zones using the Commuting Zone Crosswalk.

- We aggregate employment and wage and salary income to the commuting zone level by sex-age-education-race-foreign born-sector demographic cells. We condense some of the demographic cells as follows:
    - Drop any observations that correspond to the population living in institutions.

- Drop any observations younger than 16 and older than 64, and group ages: 18-24, 25-39, 40-54, and 55-64.
- Group on energy-industry sector: for each sector, industries in that sector are listed with their ind1990 code first.
    - Extraction
        - 42 - Oil and gas extraction; includes Support activities for mining
        - 41 - Coal mining
    - Generation and Distribution
        - 450 - Electric light and power
        - 452 - Electric and gas, and other combinations
        - 451 - Gas and steam supply systems
        - 422 - Pipe lines, except natural gas
    - Refining and Wholesale
        - 200 - Petroleum refining
        - 201 - Miscellaneous petroleum and coal products
        - 552 - Petroleum products
    - Energy-intensive manufacturing (defined as manufacturing industries above the 90th percentile of average energy used per dollar of shipments for the 2014-2018 period using the NBER-CES Manufacturing Industry Database)
        - 142 - Yarn, thread, and fabric mills
        - 160 - Pulp, paper, and paperboard mills
        - 192 - Industrial and miscellaneous chemicals
        - 252 - Structural clay products
        - 250 - Glass and glass products
        - 251 - Cement, concrete, gypsum, and plaster products
        - 270 - Blast furnaces, steelworks, rolling and finishing mills
        - 272 - Primary aluminum industries
    - Automobile sales and services
        - 612 - Motor vehicle dealers
        - 620 - Auto and home supply stores
        - 751 - Automotive repair and related services
        - 750 - Automobile parking and carwashes
    - Automobile manufacturing
        - 210 Tires and inner tubes
        - 351 Motor vehicles and motor vehicle equipment
        - 500 Motor vehicles and equipment (wholesale)
    - Gasoline service stations.
        - 621 Gasoline service stations

    - Group education: Less than 4 Years of College ("educ" less than or equal to 100) and 4+ years of college ("educ" in 101-116).
    - Group on race: non-Hispanic white (race = 1 & hispan = 0) and other (all else).
    - Group on foreign-born status: foreign-born (born outside United States or outlying areas/territories, bpl > 120; exclude those born abroad to American parents, citizen > 1) and native born (all else).
    - Create employment using values of "empstat" = 1 (i.e. individuals reporting that they are employed).
    - Create annual weeks worked using the midpoint of the range of weeks worked in "wkswork2".
    - Create annual hours worked by multiplying our weeks worked variable by the usual "uhrswork".
    - Create a total income earned variable by summing wage and salary income with business and farm income.
    - Crosswalk all occupation and industry codes using the Occupational Code Crosswalk and Industry Crosswalk.
- We append the aggregated files for each year (1980, 1990, 2000, 2010, 2011, 2017, 2022) together.
- We then construct the following variables:
    - ***ftmedinc_wage:*** the median yearly wage of people employed in the sector in a given czone and population subgroup, restricted only to people who were working full-time at the time of the survey
    - ***emp*** **:** number of people employed in the sector at time of survey

- Finally, we crosswalk in each state that a commuting zone belongs to for clarity in situations when a commuting zone spans across multiple states.
- The final product contains the below list of constructed variables:
    - ***year***
    - ***sex***
    - ***age***
    - ***ed***
    - ***white***
    - ***foreign***
    - ***czone***
    - ***cz_name***
    - ***sector***
    - ***emp*** **:** Employment status, not in armed forces
    - ***ftemp*** : Employment status, not in armed forces, 30+ hours/wk and 40+ wks/yr
    - ***ftmedinc_wage:*** Median wage & salary income, full-time workers

2.2 Renewables potential

- NREL developed a model called Renewable Energy Potential (reV) to estimate the annual technical generation potential (in MWh) of each county in the contiguous USA. Their database has this measure for different types of generation technology, but for our purposes, we use land-based wind and utility-scale solar photovoltaics only.
- The model starts with resource data (wind speed, tilt, azimuth, etc.) to estimate capacity factor, which is combined with modeled site-based levelized cost of energy. It then calibrates land availability by running a spatial exclusion module that handles protected areas (eg. DoD lands), urbanized areas, natural features, and terrain features. Finally, reV incorporates technical potential with data on transmission grids to calculate the supply curve using a spatial sorting algorithm. The resulting numbers from the pipeline represent the theoretical generation potential of each county for each type of energy source.
- The county-level data is then merged to 1990 commuting zones (CZ) using David Dorn's crosswalk (which we modified to reflect more recent changes in county boundaries).
- To get CZ potential generation, county potential generation values (MWh) are summed across counties in a CZ. These are then divided by CZ land area to calculate annual technical generation potential per square km at the CZ level (MWh/km^2).

2.3 Electricity generation

Data Preparation

- **Data Scope:**
    - Includes generator data from EIA Form 860, covering the years 1998–2022. Only 2001-onward is currently included in our public dataset.
    - Aggregated to the **county** and **commuting zone (CZ)** levels.
- **Data Cleaning Steps:**
    - **Status:**
        - Updated to two broad categories:
            1. **Operating**: Generators that are actively in use.
            2. **Other**: Includes back-up, standby, and other non-operational statuses.
        - Only operating generators are included in our public dataset
    - **Energy Source:**
        - Cleaned to identify broad categories (e.g., Fossil Fuel, Renewable Fuel) and detailed subcategories (e.g., Coal, Biomass, Wind).
    - **Location Adjustments:**
        - **County Name and County FIPS**:
            1. County names cleaned and matched to FIPS codes using a crosswalk.
            2. Missing county names imputed based on available county codes.
        - **State Name and FIPS**:
            1. Inconsistencies between ZIP codes and state names corrected.
    - **Nameplate Capacity:**
        - Missing values imputed by:

        1. Using values from previous or subsequent years.
        2. Converting kilowatts to megawatts for pre-2001 data.
    - Summed for aggregation to county and commuting zone levels.
- **Data Aggregation:**
    - Aggregated generator data to the county and commuting zone levels.
    - Metrics include:
        - **Plant Count**: Total number of plants per location.
        - **Generator Count**: Total number of generators per location.
        - **Nameplate Capacity**: Total generation capacity in megawatts per location.
- **Dropped Observations:**
    - Generators with no operational capability (e.g., planned, indefinitely postponed, canceled).
    - Observations missing **County**, **County Code**, and **ZIP Code**.
- **Output dataset contains the following variables:**
    - ***npcap***: Maximum output capacity of a generator under specific conditions (in megawatts, imputed where missing).
    - ***status*****:** Operational status of the generator (e.g., Operating, Retired, Standby).
    - ***energy1_broad***: Primary fuel or energy source used by the generator (e.g. Fossil Fuel, Renewable Energy)
    - ***energy***: Detailed classification of energy sources (e.g., Coal, Nuclear, Biomass, Solar, Wind).
    - ***plant_ct***: Total number of plants in commuting zone-year-energy source pair.
    - ***gen_ct***: Total number of generators in commuting zone-year-energy source pair.
    - ***czone***: Commuting zone (CZ) associated with a county, derived from Dorn's County FIPS code-CZ crosswalk.
    - ***year:*** Year of measurement

3 Data Comments

3.1 Top-coding

Age (in years) and all income variables (in base dollar amounts) are subject to top-coding, whereby values above a certain threshold are replaced using the state mean of all cases greater than or equal to this threshold value. This threshold is the 99.5th percentile in each state for the ACS from 2003 onward, $200,000 for 2000, $140,000 for 1990 and $75,000 for 1980 (see further information from IPUMS).