California Demand Response Potential Study Phase 4: Cluster load shape data user guide

Author and contacts: Sarah J. Smith (sjsmith@lbl.gov)
Samanvitha Murthy (SamanvithaMurthy@lbl.gov)
Date of dataset completion: June 2022
User guide last updated: June 25, 2024

1 Overview

This dataset contains electricity demand profiles for clusters of customers across California's three Investor-Owned Utilities (IOUs). The data is an output of Phase 4 of the California Demand Response (DR) Potential Study, and serves as the foundational baseline load model for the study's estimates of DR potential. More information about the DR potential studies, including a link to the Phase 4 report and its appendices, can be found at https://buildings.lbl.gov/potential-studies . Users of this data are strongly encouraged to read the report sections pertaining to the LBNL-Load model, the module that generates this data, including the entirety of Appendix A: Load Modeling Methodology.

The purpose of this document is to provide a description of the file structure and data taxonomy to users of the dataset. Some methodological considerations may be included, where they are likely to impact basic processing and analysis of the data. However, users should refer to the Phase 4 report and appendices for the majority of important information including descriptions of certain field variables, key assumptions, and limitations.

File structure and scenarios

The data is structured with **one folder for each scenario**. In the publicly released dataset, the only scenario variations are weather year ("1in2" or "1in10", or "actual" for 2019 only) and year (2019, 2025, 2030, 2040, and 2050). The scenario/folder names have additional flags that are unlikely to be relevant to users. For completeness, these are:

- anonymization type: only "anonymized" is released
- IEPR forecast level: only "MidDemand" is used in the study
- forecast step: "with_EE_FS", referring to "with energy efficiency and fuel substitution" is the final step in the model, and only this step is released for forecast years; for 2019, the year of the meter data used in the model, no adjustments are made and the corresponding flag is therefore "no_EE_FS"

Within each folder, there is a **"cluster summary" file** (cluster_summary.csv) that gives descriptive information about each customer cluster in the dataset for that scenario. While cluster names and demographics are consistent across scenarios, total consumption values, customer counts, and end use penetrations will vary, so it is important to always use the cluster summary file for the specific scenario you are working with. The taxonomy for this file is in the following section.

Additionally, there is **one .csv file for each cluster**, which contains the hourly electricity demand by end use. A taxonomy of those files is provided in the third section.

Finally, we have provided files showing the CAISO-wide gross and net load totals we have forecasted for the study. More details on that data are in the final section of this guide.

2 Cluster summary taxonomy

Each scenario contains a cluster_summary file that gives descriptive information about each customer cluster in the dataset. The "Demographic data" and "Clustering" sections of Appendix A of the Phase 4 report describe in detail the categories and possible values for clustering and the process for forming and labeling the clusters. Users of this data are encouraged to read through those sections in detail.

The columns in the data are as follows:

- **name**
  Cluster name. Structured as sector-util-building_type-size-climate-care-lca-lshp-kwh_bin. There is a single unique row for each cluster in each scenario's cluster_summary file.

- **kwh_bin**
  Percentile of annual electricity consumption relative to all customers that are in the same "cluster group", where a cluster group would have the same cluster name except for the final kwh_bin flag. For example, the two clusters
  res-sdge-mult_fam-allSize-marine-nonCare-NoLCA-Flat-0_50pct
  and
  res-sdge-mult_fam-allSize-marine-nonCare-NoLCA-Flat-50_100pct
  Were formed by first grouping all of the customers that fit the criteria res-sdge-mult_fam-allSize-marine-nonCare-NoLCA-Flat, and then splitting them in half based on their annual kWh demand; such that the half with less demand end up in the 0_50pct bin, and the half with more demand end up in the 50_100pct bin. The goal is always to split the group into quintiles, but fewer groups are used when there are not enough customers available. In some cases, the group is not split at all, and the kwh_bin is 0_100pct.

- **sector**
  Residential (res), commercial (com), industrial (ind), agriculture (ag), or other (other). See page 7 of the report appendix.

- **util**
  IOU (utility) service area in which the cluster customers reside. Either PG&E (pge), SCE (sce), or SDG&E (sdge).

- **building_type**
  In the residential sector, building type is single family (sin_fam), multi-family (mult_fam), master meter (master_mtr), or unknown (res_unknown). PG&E data did not contain residential building type flags, so all PG&E residential clusters have the res_unknown designation. Multi family refers to a single residence metered individually, while master meter could either refer to multiple units metered together, or to common spaces in a residential building. In the commercial sector, refers to the building type as determined by NAICS code. In the industrial sector, refers to the industrial subsector as determined by NAICS code. In the agricultural sector, refers to the type of agricultural activity as determined by NAICS code. See Table A-1 in the report appendix.

- **size**
  Refers to the electrical demand size of the customers within the cluster, based on the peak demand. Determined by the customer's actual peak demand as reported by the utility, or, if not available, by the tariff the customer was on. Note that size categories are consistent across IOUs in this study, not unique to the size categories of each IOU's tariffs. In this category, there could be combinations of indicators (e.g. med+large) or clusters could be a combination of all sizes (allSize); additional columns give the fractional breakdown across size categories. All residential clusters have the allSize distinction. See Table A-2 in the report appendix.

- **climate**
  Climate region (marine, hot_dry, or cold) of the customers in the cluster. Determined by the geographical location of the customers. Climate regions are groups of CEC building climate zones. Note that there could be combinations or "allClimate" in this category; additional columns provide the fractional breakdown of customers across climates. See Table A-4 in the report appendix.

- **care**
  CARE[1] status of the customers in the cluster; Care or nonCare. Only used for residential sector, all other sectors (and perhaps some residential clusters) have an "allCare" flag.

- **lca**
  Local Capacity Area. Determined by the geographical location of the customer. Note that many regions of the IOU's service territories are not designated as an LCA (NoLCA). Note

[1] https://www.cpuc.ca.gov/consumer-support/financial-assistance-savings-and-discounts/california-alternate-rates-for-energy

that there could be combinations or "allLCA" in this category; additional columns provide the fractional breakdown of customers across LCAs.

- **lshp**
  Load shape cluster, used for residential and commercial sectors. Groups customers based on their typical daily patterns - where the lshp describes a time of day (e.g. EarlyEve), this describes when those customers' load tends to peak. Note that there could be combinations or "allLshp" in this category. See Figures A-9 and A-11 in the report appendix.

- **ts_count**
  Number of time series (AMI data files) that were used when building the cluster.

- **customer_count**
  Number of customers the cluster represents.

- **kwh_ann_gross**
  Total kWh consumed by the cluster customers' end use demands in a year.

- **kwh_ann_net**
  Net kWh served by the grid for the cluster customers. kwh_ann_gross less total behind the meter PV generation.

- **care_frac**
  Fraction of customers in the cluster that are on a CARE[1] rate. Binary except for "allCare" clusters.

- **cca_frac**
  Fraction of customers in the cluster that are Community Choice Aggregation customers.

- **da_frac**
  Fraction of customers in the cluster that are direct access customers.

- **all_elec_frac**
  Fraction of customers in the cluster that are flagged as being all-electric customers (no gas service).

- **pv_penetration**
  Fraction of customers that have PV panels.

- **dac_frac**
  Fraction of customers in the cluster that are located in a disadvantaged community.

- **{climate}_frac**
  Fraction of customers in the cluster that are in the given climate region (hot, med_dry, or cold).

- **{size}_frac**
  Fraction of customers that are in the given size category (large, med, or small).

- **{lca}_frac**
  Fraction of customers in the cluster that are in the given local capacity area (NoLCA, LA, Ventura, Bay, North, Humboldt, Sierra, Stockton, Fresno, Kern).

- **{enduse}_penetration**
  Fraction of customers that have the given electrically-powered end use (cooling, heating, water_heating, solar_water_heating, cooking, dishwasher, spa_heater, spa_pump, washer, dryer, outdoor_lighting, television, office_equipment, pc, pool_pump, ventilation, other, refrigeration, freezer, indoor_lighting, ev_level1, ev_level2, it_equipment, boiler, proc_heat, proc_cool, proc_mach, proc_elch, proc_other, nonproc, pumping).

- **EV_Count**
  Number of EVs estimated to charge in the cluster. For com clusters in particular, this is more informative than the penetration.

- **{MHDEV type}_count**
  For MHDEV clusters; number of the given vehicle type (medium_duty_truck, agriculture_truck, other_freight_truck, construction_truck, utility_truck, tractor_trailer_truck, drayage_truck, refuse_truck, bus).

3 Time series taxonomy

Each cluster's time series file contains the modeled hourly electricity demand, broken out by end use, for that cluster.

It is important to note that the calendar structure of the time series does not vary by forecast year. The analysis was based on 2019 data, and all time series files maintain that calendar structure, meaning that all time series files start on a Tuesday, with the 5th and 6th days of each file (and every 7 days after) representing weekend usage patterns.

All cluster time series files contain the following columns

- **hour_ending**
  The hour of the year for the corresponding load. 12:00-1:00 AM on January 1st is considered hour_ending 1, and the electricity demand for each subsequent hour of the year is labeled accordingly.

While time of day is not explicitly given in the time series files, the data can be considered to represent Pacific Standard Time. For example: if a customer starts using electricity at exactly 8:00am every day, according to prevailing time (the time experienced by people after accounting for daylight savings time), the load would increase every 24-hour period for the first few months of the year. Then, when daylight savings time occurs, the load would increase only 23 hours after the prior increase, since the "8:00am" clock hour moved forward one hour. The 8:00am spike would then continue to occur every 24 hours, until the Fall, when one 8:00am spike would come 25 hours after the prior spike as the clocks are moved back. For reference, 2019 Daylight Savings time began on March 10th, when 2:00AM became 3:00AM. This corresponds to hour_ending 1635. It ended on November 3rd, when 2:00AM became 1:00AM. This corresponds to hour_ending 7347.

- **total**
  Gross electricity demand of all of the end uses in the cluster in each hour, in units of kWh.

- **pv_generation**
  Total behind-the-meter PV generation for the cluster in each hour, in units of kWh. The electricity demand of the cluster on the grid can be calculated by subtracting the pv_generation from the total. Note that the pv_generation may be larger than the total demand at times, implying that the cluster is exporting to the grid.

Additional columns represent the electricity demand of each end use. These columns vary primarily by sector, with some columns missing from certain cluster time series files if the given end use is not present in the cluster. See Table 2 in the Phase 4 report for a full list of end uses.

4 Syste load totals

The zip file gross_net_load_totals.zip contains the modeled CAISO-wide load totals used in this study. There are separate files for each load type (gross or net) and each weather year, with each file containing a column for each year in the analysis. The **gross load** represents the total demand on the grid, net of behind-the-meter PV generation, while the **net load** represents the gross load less front-of-the-meter variable renewable energy (VRE; i.e., wind and solar PV) generation.

Because the IOUs do not constitute the entirety of the CAISO grid, we scale up the total demand modeled across all customer clusters in order to perform system-level analyses. This scaling factor is 1/0.798, as we estimate that our customer load clusters do not account for 20.2% of CAISO demand. In other words, when VRE generation is zero, the sum of demand across all the cluster time series files will be equal to the gross load total file multiplied by 0.798. Our estimates for system-level VRE generation can be backed out by subtracting the net load total from the gross load total. Lastly, it should be noted that grid-scale battery storage is not accounted for in this model. Therefore, particularly in future forecast years, significant periods of overgeneration are shown in our load totals.