FILED
03/30/23
03:22 PM
A2205002

ATTACHMENT 1

Development and use of cluster load shapes for the Phase 4 DR Potential Study

1 Introduction

This document briefly describes:

1. The purpose for publicly releasing the anonymized customer cluster load shape data developed for the Phase 4 California Demand Response (DR) Potential Study conducted by Lawrence Berkeley National Laboratory (LBNL)
2. A summary of related work products to be released (pg. 2)
3. The technical procedure for developing the anonymized customer cluster load shapes (pg. 2-5)
4. The technical procedure for use of the cluster load shapes in modeling DR potential and generating DR supply curves (pg. 5-8)
5. The details of the work products to be released, including the presentation of customer cluster load shape data (if any) in each product. (pg. 8)

The release of the anonymized customer cluster load shape data is intended to provide transparency as to the data underpinning the Phase 4 Potential Study. Given the level of detail and granularity with which they represent customer classes and end uses, these cluster load shapes also have the potential to provide significant value in future modeling and analysis efforts by other organizations within the state. The effort by LBNL to produce these customer cluster load shape data was considerable, including significant data processing, modeling, and forecasting efforts. Making the data product public in support of future analyses will therefore enhance the value of this effort to ratepayers. Possible future use-cases for public cluster load-shape data include, among others:

1. Demand forecasting for the CEC Integrated Energy Policy Report (IEPR) and CPUC Integrated Resource Planning (IRP) proceeding, including the impacts of electrification on system load shapes
2. Informing IOU delivery-system planning, and IOU and CCA resource-adequacy planning, by providing detailed load shapes and forecasts for loads with high growth potential, such as electrified loads
3. Modeling the load-flexibility capacity of emerging technologies and determining their potential for future research and development to support California demand flexibility needs, by regulators, researchers, or technology firms
4. Modeling the potential customer response to future approaches to achieving demand flexibility, including in the context of the CPUC's current dynamic pricing proceeding
5. Developing targeting strategies for engaging customers in future demand flexibility programs, by IOUs, CCAs, program administrators, or third-party aggregators
6. Planning future energy efficiency and demand response programs, by utilities, regulators, or third-party aggregators
7. Improving future estimation of savings potential, avoided costs, and cost-effectiveness for energy efficiency and demand response measures, by regulators, IOUs, aggregators, or technology firms

8. Improving calibration of building energy models such as those used to determine energy savings of EE retrofit projects by contractors
9. Enabling intervenors in CPUC, CEC, or CAISO policy processes and proceedings to perform analyses of load impacts, emissions, bill impacts, etc., based on high quality, realistic data representing California customer electrical loads

2 Work products for release – Summary

The anonymized customer cluster load shape data and results derived therefrom are intended to be released in several public work products as summarized in the table below and described further in the section at the end of this document. For purposes of transparency, LBNL will also release the modeling software used for the Phase 4 DR Potential Study, which includes a load modeling module ("LBNL-Load") to develop the cluster load shapes, as well as a DR potential estimation module ("DR-Path") for which the load shapes are used as inputs.

Work product for public release	Level of data aggregation used to prepare work product	Data products released to the public
Phase 4 California Demand Response Potential Study	Partially anonymized data: ● 15+ non-residential customers per cluster ● 100+ residential customers per cluster ● 8760 load shapes ● Minimum LCA-level geographic aggregation	Tables and graphs showing: ● Seasonally averaged cluster load shapes ● Aggregated system-level load shapes ● Aggregated DR potential derived from analysis of the cluster load shapes
Detailed demand response potential results	Partially anonymized data as described above	Data files describing the cluster- and end-use level DR potential results
Anonymized cluster load shape data release	Fully anonymized data: ● As above, plus no more than 15% of any cluster's total load deriving from any one customer	One data table for each anonymized cluster in machine-readable (CSV) format, corresponding to the full 8760 load shape for the total cluster load and each disaggregated end use
Prototypical daily load shapes	Prototypical load shapes are generated for each sector. Information about the frequency of load shape occurrence may be shared at a more granular level, but never for a group of less than 30 customers.	Graphs and data files describing the prototypical load shapes (i.e. load shape clusters) for each sector, and the fractional breakdown of prototypical load shapes that comprise a given group of customers' usage patterns.
Customer enrollment model	Regression models are specified for each sector, building type, building size, CARE status, and climate region.	Fractional regression models that estimate the portion of customers that will enroll in a DR program for a given incentive level.

Phase 4 DR Potential Study modeling software modules	No customer data used in this work product	LBNL python source code for Phase 4 modeling
Phase 4 DR technology cost/performance data	No customer data used in this work product; based on secondary data in the published literature	Tables of technology load impact and cost, for specific DR technologies

3 Customer cluster load shape development—Technical Methodology

3.1 Customer data

LBNL completed a 2-stage IOU data request in Spring of 2020. The first stage obtained demographic and 2018-2019 annual electricity consumption data on all ~13 million IOU customer accounts that were active in 2019. From this information, LBNL developed a stratified random sample of 3% of accounts for which interval meter data was requested. As part of the sampling process, each sampled customer was assigned two weighting factors:

- A customer weight, indicating the total number of customers the sampled customer represents within its sampling group
- An energy weight, indicating the total amount of annual energy consumption the sampled customer represents within its sampling group.

Values for these weights range from 1 to more than 1000, depending on the customer segment being sampled. In total, interval meter data for 411,000 meters were collected for 2018 and 2019.

The meter data may be on a 15-minute or hourly basis, and are separated into energy delivered to the customer and energy received from the customer (for customers with behind-the-meter generation). The relevant steps to processing these data into individual time series are as follows:

- 15-minute data is aggregated to the hourly level
- Data from delivered and received channels are combined to create a net demand profile
- Data is corrected for Public Safety Power Shut-off events (during which no data are available), with simulated data being filled in for the relevant time periods.

3.2 Customer clustering

A statistical clustering algorithm is applied to the customer load shapes to subdivide them into load shape clusters that represent a set of prototypical load shapes, allowing customers to be grouped according to shared behavior patterns. The customers are then further segmented into more fine-grained customer clusters by subdividing the customers along the following dimensions:

- Sector
- Utility
- Building type

- Customer size category[1]
- Climate region[2]
- Receipt of CARE subsidy
- Local capacity area (LCA)
- Load shape cluster
- Total annual energy consumption.

In this procedure, the smallest geographic region on which customers are aggregated is an LCA, which is much larger than the ZIP code level that is typically used as the minimum in customer anonymization criteria.

The segmentation proceeded through this list of characteristics using a hierarchical strategy. First, customers were subdivided by sector, then utility, then building type, and so on until the list of characteristics was completed, or until it was impossible to create a cluster containing at least a preset minimum number of customers, N_{min} (calculated as the sum of customer weights of the time series samples). Where no segmentation was possible on a given characteristic (e.g., load shape cluster in the industrial and ag sectors), that level of the hierarchy was skipped. If segmenting on a particular characteristic would yield fewer than N_{min} customers in a cluster, we recursively recombined clusters with neighboring clusters until we obtained a sufficiently large cluster. For the clustering in this study, we set $N_{min} = 100$ for residential clusters and $N_{min} = 15$ for non-residential clusters, to support the anonymization step below.[3] This procedure yielded a set of 5422 customer clusters across all sectors and IOUs.

3.3 Cluster load shape aggregation and anonymization

Finally, for each cluster, we aggregated the 2019 load data of all sampled customers that belonged to the cluster, weighted by their assigned energy use weights from the sampling process, to produce an hourly cluster-level aggregated load shape for each cluster. For residential customers, we also applied load shape adjustments to account for the transition to default time-of-use tariffs after 2019, based on IOU-published load impact reports. The resulting aggregated cluster load shapes represent the primary load inputs to the DR-Path model in the Phase 4 study.

We also computed a separate set of cluster load shapes that were further processed to adhere to the "15-in-15" anonymization criteria that are often used for the public release of energy consumption data. This requires that each cluster represent consumption from at least 15 customers (100 in the residential sector), with no customer representing more than 15% of the total consumption in the cluster. Since we set N_{min} appropriately when creating the clusters, the first criterion was met. To meet the second criterion, for any

[1] Customers were subdivided into approximate small, medium, and large subcategories according to their peak demand as reported in the IOU data.

[2] In this study, CEC Title 24 climate zones were mapped to three aggregate climate regions (hot-dry, marine, and cold) in accordance with EE Potential and Goals Study.

[3] In addition to these limits on the total number of customers in each cluster, we also required that each cluster contain at least 15 time series samples to ensure a reasonable statistical sample in each cluster.

cluster with more than 15% of consumption from a single customer, we adjusted the customer weights recursively and redistributed the surplus weight to other members in the cluster until no customer represented more than 15% of the total cluster consumption. These anonymized clusters were not used in the study but they are intended to be released publicly as an approximate, fully anonymized representation of the study input data.

The cluster load shapes are calculated as a weighted sum of customer time series, which is then reweighted to meet the anonymization criteria. No information will be released publicly regarding the number of time series used to create a cluster's load shape, the customers whose time series were sampled/used, the weights used in the aggregation, or any other information describing the specific customer load shapes that have been aggregated. There is no way to extract a single customer's time series profile based on the cluster's load shape and other information included in the cluster load shape dataset.

3.4 End use disaggregation

The aggregated cluster load shapes represent the total hourly load from all customers in each cluster. These were then disaggregated into different individual electrical end uses using statistical estimation procedures developed for the Phase 4 study modeling. Each resulting end use profile represents the total hourly consumption deriving from a given end use (e.g., refrigeration or lighting) from all customers within a cluster. This disaggregation does not add any information that could be used to identify individual customer loads within the cluster.

3.5 Final cluster load shape data

The result of the procedure above is two sets of 5422 cluster load shapes, one that has been fully anonymized according to the 15-in-15 rule, and one that has not. Each load shape consists of several 8760 hourly time series representing the estimated total load from all customers in the cluster, within a defined set of end uses and in total.

Because they more accurately reflect total load, the non-anonymized cluster load shapes were used in the Phase 4 study, but they are not intended for public release. The anonymized cluster load shapes are intended for public release, pending CPUC approval, to provide an approximate representation of the study input data. Similarly anonymized cluster load shapes were released in concert with the Phase 2 and Phase 3 DR Potential Study.

The modeling code that was used to develop these load profiles (but not the underlying data inputs) may also be released for purposes of transparency, as was also done in previous phases of the study. As discussed above, it is not possible to extract individual customer load shapes from the anonymized cluster load shapes without detailed knowledge of the underlying customer weights and anonymization reweighting values used for aggregation. This is true even if the precise analytical methods are known. Data on the sampling and anonymization weights will remain confidential; hence **it will not be possible to extract individual customer load shapes from the public data or code**.

4 DR potential and supply curve modeling – Technical methodology

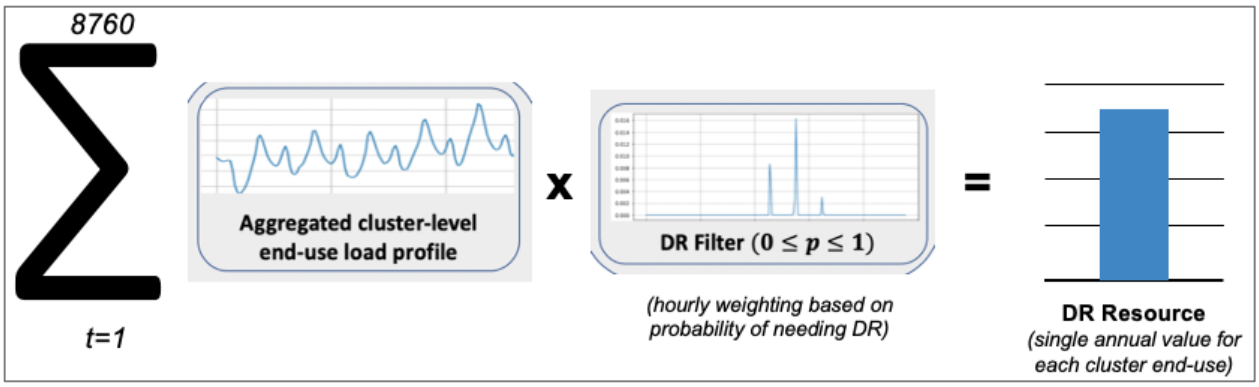
As described above, the non-anonymized load shapes are not intended for public release, but they were used in calculating DR potential results that are intended for release via the Phase 4 report and other ad-hoc requests for study results. Therefore, this section describes in detail how those load shapes were used, and to what extent the resulting DR potential estimates are sufficiently removed from them such that no time series information can be discerned from them. The overall modeling pipeline consists of the following high-level steps:

1. **Cluster load shape modeling**, which uses customer meter data and other resources to generate cluster-level load shapes, disaggregated at the end use level
2. **DR resource calculation**, which uses hourly probabilities of DR dispatch to estimate the average annual DR resource available from each cluster end use
3. **Generation of all possible DR pathways**, where each cluster end use is combined with all relevant enabling technologies and various incentive levels to calculate all possible DR "pathways", defined by a quantity of DR potential and associated cost.
4. **Creation of supply curves from DR pathways**, which generates DR potential supply curves by intelligently aggregating the DR-Path results database consisting of all possible DR pathways at various price levels.

Step 1 is described in detail in the above section of this document. Here, we will describe steps 2-4, which collectively make up the DR-Path model, with a focus on how the cluster load shapes are used. **Step 2 is of particular relevance, as it is when the temporal load shape data is "lost" via calculation of annual weighted averages; after this point, no intermediate or final results provide temporally-specific electricity demand information.**

4.1 Calculating a "DR resource" metric for all cluster end uses

The DR Potential Study reports DR potential on an annual basis, with the units of kW (for Shed DR) or kWh (for Shift DR) representing the average DR resource over the course of the year. However, the need for DR varies significantly throughout the year, so instead of taking a simple annual average resource we calculate a weighted average resource, where each hour is weighted based on how likely it is that the grid will need DR. In the model, we first calculate this weighted average annual resource for each cluster end use based purely on the load profile, before technology capabilities or enrollment is considered. The result is referred to in this document as the "DR resource"; the calculation of this metric is shown in the figure below. As shown, the hourly demand profile for each end use in each cluster is multiplied by a "DR filter", which is an hourly weight describing the likely need for DR. These values are then summed across the 8,760 hours in the year to create the DR resource metric.



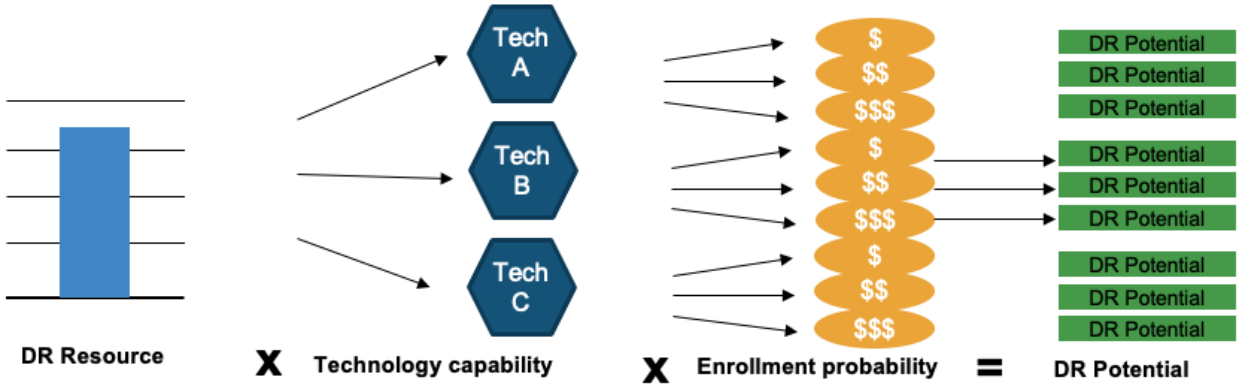
There are two critical details here that should be emphasized when considering data privacy:

1. The aggregation over the year **removes all temporal information** of the cluster load shape
2. The DR Filter skews the demand in each hour such that **even the aggregate annual load** of a given cluster **is not revealed** via the DR resource metric.

Separate DR filters are used for Shed and Shift DR, since the need for Shed and Shift occurs at different times of the year. Therefore, a DR resource metric is calculated for every cluster, for every end use, for each type of DR being considered.

4.2 Estimating DR potential for all possible pathways

The core of the DR-Path involves taking the DR resource metrics described above and determining what the cost and ultimate potential would be if that resource was enabled to perform DR via various technology and incentive pathways. An illustrative example of numerous pathways for a single DR resource is shown in the figure below. As shown, each DR resource (representing a single end use for a single cluster for a given type of DR) is paired with each possible enabling technology, with the resource metric being multiplied by parameters that capture load response capability of the given technology. Then, each technology's controllable resource is paired with numerous possible incentive levels that impact what portion of the customer base is likely to enroll in DR. The resulting metric is the "DR Potential" of the pathway; the cost of procuring this pathway is also calculated in parallel.



These DR pathways are stored in a DR-Path results database, which defines for each type of DR (Shed and Shift), for every cluster, for every end use, for every possible enabling technology, and for every possible incentive level, what the cost and total potential of DR would be. **Because these metrics are based solely on the DR resource metrics, not any more detailed information about the underlying load, they do not expose any information regarding the cluster load shapes.**

4.3 Creating DR supply curves from database of pathways

The pathway results that comprise the DR-Path results database are largely mutually exclusive, as a single DR resource cannot be enabled by multiple technologies or at multiple incentive levels. Therefore, the final processing of these results requires thoughtful selection and aggregation of pathways in order to describe the overall cost and potential of enabling DR. This is primarily done through the creation of supply curves, which show the total DR resource across the year that can be procured at increasing price levels. Supply curves are generated separately for Shed and Shift DR, using the following basic algorithm:

- For each price level (e.g. $25-500 in $25 increments):
    - For each cluster end use:
        - Remove all DR pathways with a cost above the price level being examined
        - Of the remaining pathways, select the one with the highest DR potential
    - Aggregate the DR potential available across all cluster end uses to determine the total DR potential at the given price level

When creating these supply curves, it is often desired to show how much of the DR potential is coming from various end uses, technologies, or customer types. Therefore, when the aggregation step occurs at each level, we often aggregate within these categories and "stack" them into the final supply curve. In any case, **these supply curves are aggregations of the DR pathways, which are based on the DR resource metrics, and therefore no information about the underlying cluster load shapes is exposed**, even at the most granular level.

5 Work products for release – Details

5.1 Phase 4 DR Potential study

The non-anonymized cluster load shapes are used as inputs to the DR-Path model, which reports potential DR resources (GW of shed DR potential or GWh of shift DR potential) aggregated across various dimensions of the clusters, such as sector, IOU, building type, LCA etc, via numerous supply curves. The report will contain supply curve figures, and in some cases, tables with the associated supply curve numbers. These results do not carry any time series information about the underlying cluster load shapes, as described in the section above. Charts displaying seasonally averaged load shapes for a set of example clusters are also presented. The seasonal averaging makes it impossible to extract any detailed hourly load shape data from these charts. Additionally, the appendices to this report include the prototypical daily load shapes used to determine whether customers exhibited certain usage patterns, e.g. flat load shape,

night peaking load shape, and so on, along with the customer enrollment model. Each of these are described in the subsections below.

5.2 Detailed DR potential results

As described above, the Phase 4 report will contain DR Potential results in the form of supply curves that are generally aggregated to show results across the entire state by sector, end use, IOU, or some other indicator. Additionally, there have been requests to release the more detailed DR potential results to third parties to utilize these numbers in their own analyses, allowing them to filter and aggregate the results according to their needs. This data would essentially provide the most granular supply curve possible, with a single datapoint for each cluster and end use combination in the study. As described in section 4 of this document, these results stem from the "DR resource" metric that is a scaled and aggregated transformation of the cluster time series. Therefore, even at this most granular level, no temporal or total usage information about any given cluster is exposed.

5.3 Anonymized cluster load shape data release

In addition to presenting modeling results based on the cluster data, the anonymized cluster load shapes will also be released as a public resource. The dataset will consist of a set of 5422 text files in CSV format (one for each cluster) containing tables representing 8760-hour time series of the load from each modeled end use for each cluster. Released data will adhere to the anonymization criteria described above.

The release will provide transparency into the input data of the study. The anonymized cluster load shape data will also serve as an important resource for future research and modeling into demand patterns and growth in California on the level of detailed customer types. For instance, the data can support future demand forecasting by the CPUC, CEC, or others. It can also support investigations into the demand-side resources (either energy efficiency or demand response) that are available among different types of customers, enabling more effective targeting of customer classes who may have significant potential.

5.4 Prototypical daily load shapes

As described in Section 3.2, one of the indicators used to group customers into clusters is the customer's "load shape cluster", which refers to a qualitative description of the customer's daily usage patterns. Examples of load shape clusters include "Flat", "LateEve", or "AllDay". These load clusters are formed starting with clustering every day of every customer's data into about 60 groups, with the average of all daily load shapes in the group representing the initial cluster. The shapes of each of these clusters, which we refer to as "prototypical daily load shapes", and the fraction of customers' days that fall in the various clusters, is of great interest to the research community. This clustering is performed on a sector-by-sector level, with tens of thousands (or, in the case of residential, hundreds of thousands) of customers' data going into the analysis. While this data describes common daily patterns of electricity demand, **the released prototypical daily load shapes are averages of hundreds of thousands of daily load profiles and do not relate back to any customer, or any small group of customers.**

Information on the *prevalence* of each prototypical daily load shape has been requested for more granular sets of customers, such as by building type and size. To answer such requests, we may say "70% of daily

load profiles for large office buildings follow a *Flat* pattern, while 30% follow an *AllDay* pattern", for example. No information regarding the prevalence of daily load shapes will be released for a group of fewer than 30 customers. Further, describing the prevalence of daily load shapes does not directly release an aggregated or averaged group of load shapes; as we only indicate the prevalence of the load shapes generated at the whole-sector level.

5.5 Customer enrollment model

The customer enrollment model will be shared with third parties on an as-requested basis, and may be part of the Phase 4 modeling code released (as described below) if approved. This model was built using fully anonymized SCE residential, commercial, and industrial customers' data on demand response program enrollment data. However, none of this data is present in the resulting regression equations that form the "customer enrollment model". These regression equations simply describe an approximation of the aggregate relationship between incentive level and enrollment for each customer group**; they do not describe the specific programs that customers did or did not enroll in, how many customers enrolled in programs, or any other customer information**.

5.6 Phase 4 modeling code

For purposes of transparency, LBNL will also release its modeling code that was used to develop the customer cluster load shapes, as well as LBNL's code used to process the load shapes into estimates of DR potential. This code will provide insight into the methods used to develop the clusters and to process them into DR potential estimates. It will not be possible to meaningfully run the load-shape modeling code using only the public data; the release is purely for purposes of methodological transparency. In particular, none of the customer or cluster load shape data described here is included or embedded in the modeling code. It is not possible to use this code to reverse-engineer individual customer information from the aggregated load shapes. As described above, doing so would require access to data inputs that will not be made public in any of the work products described here.

5.7 Phase 4 DR Technology Cost/Performance Data

For purposes of transparency, LBNL will also release the cost and performance data for each of the DR technologies included in the Phase 4 study. These data are derived from published literature and do not include any customer-specific data. These data cannot be used to de-anonymize any of the data sets listed above.

(END OF ATTACHMENT 1)