EXH. CD-7
DOCKETS UE-22___/UG-22___
2022 PSE GENERAL RATE CASE
WITNESS: DR. CHHANDITA DAS

BEFORE THE
WASHINGTON UTILITIES AND TRANSPORTATION COMMISSION

WASHINGTON UTILITIES AND TRANSPORTATION COMMISSION, **Complainant,** **v.** **PUGET SOUND ENERGY,** **Respondent.**	**Docket UE-22____** **Docket UG-22____**

SIXTH EXHIBIT (NONCONFIDENTIAL) TO THE PREFILED DIRECT TESTIMONY OF

DR. CHHANDITA DAS

ON BEHALF OF PUGET SOUND ENERGY

JANUARY 31, 2022

LOAD RESEARCH SOFTWARE (LRS)

Load Research Software or LRS is a software tool developed by DNV. DNV is an International Energy Consulting Firm with services across the utility landscape. With over 40 years of direct experience in the utility industry, DNV is a recognized international leader in the fields of energy program research & evaluation and load research & analytics. The LRS was designed for load research, but it is also suitable for program evaluation and sample design. Because LRS was designed to accommodate a wide range of application, it is undeniably complex. The LRS system is very complex and consisted of multiple steps to achieve different objectives. LRS can be used for different types of data such as MV-90, and any level of granularity such as hourly, 15-minutes or 5-minutes. The LRS system is designed to meet most of the load research needs and since it is programmed in SAS, it can be customized any specialized requirements. This section describes the two main steps in the load research analyses, data validation and sample post-stratification methodology. This description is taken from the DNV Software Documentation.

1 VALIDATION, EDITING, AND ESTIMATION (VEE) APPROACH

This section highlights the validation, editing, and estimation (VEE) approaches used within the LRS framework. The VEE process covers each individual site's load data and how to validate the load data that exists, identify bad data for editing, and estimate missing intervals with appropriate values. The major aspects of the VEE process that are covered in the DNV GL Load Research System (LRS) Software are

- Identifying suspect outages for removal;
- Identifying extreme spikes and valleys;
- Validating the load data against the associated billing data; AND
- Using weather regression models and interpolation to estimate values for missing data.

1.1 Outages

This step in the VEE process is to attempt to identify suspect outages in the load data. Suspect outages could be a result of bad metering, a failed meter, or poor meter data management estimation. Trying to identify these outages is key in ensuring that load data is being represented well. If suspect outages have been identified, the user can choose to remove this load data and attempt to estimate the outages with actual data or to remove the load profile from analysis depending on how much load data has been reset to missing.

1.2 Outlier Identification

This step in the VEE process is to attempt to identify outliers for both peaks and valleys. Outliers could be a result of bad metering, a failed meter, or poor meter data management estimation. It could also be a result of an unusual circumstance at the site's location resulting in a value that is not typical. It will be up to the user to determine if the unusual circumstance should remain for analysis purposes or if it should be removed for an expected value. Trying to identify outliers is key in ensuring that load data is being represented well, especially if the peak outlier is extreme enough to have impacts on allocators used in rate design and pricing. If outliers have been identified, the software will reset these values to missing. While one method was used prior to the VEE 2.0 update, there are now multiple techniques available for the user to implement on the load data to ensure that the intervals are truly outliers. Based on user settings, identified outliers must fail one or multiple methods before being reset to missing. The user can also manually identify intervals they wish to reset missing if they believe those intervals are inaccurate, whether they are outliers or not.

1.3 Metered to Billing Comparison

Validating the load data is an important part of the VEE process, whether it is load data that fully exists for any sample point or if estimation was used to fill in gaps. The main goal of the bill-to-meter process is to be a validation check on the load data when comparing to a known, true value. This value is often the billed usage obtained for the same sample point. The user can compare the load data on a bill-by-bill basis or at an aggregate level for the study frame being compared. Whether the user uses the actual bill read dates and days in the billing cycle to define the start and end frames for comparison or a read cycle definition to define these time frames, the process ensures that the load data being compared to the billing data are examining an appropriate time frame allowing for an "apples-to-apples" comparison.

When comparing the billing usage against the load data, the ratio between the two are compared to a defined threshold. The calculation of this comparison is:

$$1 - (RatioThreshold - 1) < \frac{\sum LoadData_i}{\sum BillData_i} < RatioThreshold$$

Where,

- *RatioThreshold* is the percent difference the load data is allowed to be within to be considered "ok". This value should be greater than 1. If the allowed threshold is 5%, then this value should be set to 1.05;
- *LoadData* is the sum of the interval usage for the given time frame *i;*
- *BillData* is the sum of the billing usage for the given time frame *i*. It could just be the billed usage as is without doing any summation; and
- *i* is the representation of a given bill or timeframe. The user could choose to sum all *i* bills for comparison or do each of them individually.

1.4 Time-Temperature Regression Modelling

Another important aspect of the VEE process is to attempt to fill in the missing data with appropriate estimated values. The importance of estimating the missing values is to attempt to have as complete load data as possible for each sample site. There will be instances that a site may have too much missing data in order to do effective estimation. The user can choose to use the data that does exist for this data or to remove the site completely from the analysis. For those that have enough non-missing validated intervals to do regression modelling, different methods can be used for estimation. Weather regression models are utilized in the modelling techniques for longer missing time frames. For shorter time frames (which can be defined by the user), interpolation is utilized to fill in those values as shorter time frames will be best represented by the non-missing intervals around it. For estimation, the best practice is to not estimate new peaks, so the user can prevent the software from creating an estimate that is greater than the non-missing max value for the sample site. In addition to these two methods, any validated sites can be used to create an average customer shape that could then be used to fill in the gaps for the sites that do not have enough load data to effectively estimate their own missing loads through regression modelling or interpolation.

This analyses uses a time-temperature modelling strategy as the fundamental basis to build models for use to fill gaps and missing data. The approach develops a mathematical model that represents the relationship between

energy usage and temperature. Using this model, intervals with missing load data can be predicted applying the temperature of that hour to the model of best fit for that hour.

This normalization analysis recognizes that each customer reacts differently to varying heating and cooling degree days, and each customer has unique space conditioning characteristics. Buildings with more efficient heating or cooling equipment, radiant barriers, more insulation, and efficient windows will consume less energy because they will require less heating and/or cooling.

The simplest model where the specifications is such that energy consumptions depends on either heating or cooling degree days only is shown in Equation 1.

Equation 1 – Basic Model

$$U_i = \alpha + \beta * DD_i(\tau) + e$$

Where;

U_i = average daily consumption in interval i.
$DD_i(\tau)$ = average degree days in interval i, based on reference temperature
α,β = parameters to be estimated to minimize e.
e = a random error term.

The base model reflects that a customer's energy usage is equal to some base level α, and a linear function between a reference temperature τ, and the outside temperature. The constant proportionality, β, represents a customer's effective heat-loss or heat-gain rate. As mentioned, the model recognizes that each customer has unique space conditioning operating characteristics. To capture these unique space conditioning characteristics, the modelling runs regressions for a range of heating and cooling reference temperatures (i.e., temperatures at which users tend to turn on heating or cooling equipment) against usage. The model chosen to represent a customer's energy use is the model that best linearizes the relationship between usage and degree days. A degree day is the difference between the recorded temperature for a period (could be 15 minute, hourly, or daily depending on how the modelling approach is being applied) and the point at which an occupant will act in response to temperature (either turning on the heating or air conditioning). For example, if a building occupant will turn on the AC at 74°F and the recorded temperature for an hour was 85°F, the total cooling degrees would be 11. A cooling degree day is the sum of cooling degrees for each day. For each customer, an optimal model based on a unique reference temperature (τ is identified by the minimum mean squared error (MSE) of the modelling regression) is selected. Models for each site are built by day of the week (DOW) and hour. Users can specify to use individual days or weekday/weekend for the model DOW.

When the model regression is applied to a customer's heating characteristics, it is referred to as the *heating only model* (HOM). When the model regression is applied to a customer's cooling characteristics, it is referred to as the *cooling only model* (COM). When the model regression model is applied to both heating and cooling characteristics, it is referred to as the *heating and cooling model* (BOTH). For this analysis, we used a BOTH model because customer can use electricity for both heating and cooling.

The analysis identifies the optimum HDD and/or CDD for each customer, which will be used to fill in gaps in the load data file using actual temperature for that DOW.

2 POST-STRATIFICATION

In order to statistically analyse the sample data, it is necessary to use one of the post-stratification modules to calculate case weights that reflect the current population. The case weight in each stratum is the ratio between the number of accounts in the population and the number of accounts in the sample with usable load data. The LRS System offers four methods for post-stratification. – Model-based stratification, balanced stratification, Dalenius-Hodges stratification, and stratification using any arbitrary cut points and allocation.

This approach follows the principals of model-based statistical sampling ("MBSS") as the basis for analysis since it is optimized for stratified ratio estimation. MBSS techniques have been used to create a very efficient and flexible structure for collecting data on countless energy efficiency evaluations, demand response evaluations, and interval load data analyses, e.g., load research and end-use metering, projects.

2.1 Background

Conventional methods are documented in standard texts such as Cochran's *Sampling Techniques*.[1] MBSS is grounded in theory of model-assisted survey sampling developed by C.E. Sarndal and others.[2] [3] MBSS methodology has been applied in load research for more than fifty years and in energy efficiency evaluation for more than thirty years. This fusion of theory and practice has led to important advances in both model-based theory and interval load data collection practice, including the use of the error ratio for preliminary sample design, the model-based methodology for efficient stratified ratio estimation, and effective methods for domains estimation.

MBSS and conventional methodologies are currently taught in the Association of Edison Illuminating Companies' *Advanced Methods in Load Research* seminar. MBSS methodology is also documented in *The California Evaluation Framework*.[4] MBSS has been used successfully for decades in countless load research and program evaluation studies. It has also been examined in public utility hearings and in at least two EPRI studies.

2.2 The Role of the Statistical Model

MBSS uses a statistical model to guide the planning and the sample design. The parameters of the model, especially the error ratio, are used to represent prior information about the population to be sampled. The model describes the nature of the variation in the relationship between any target *y variable* of the study, in our case the hourly consumption of the customer, and one or more *x variable*s that can be developed from known billing data and other supporting information. The *x variable* is usually a measure of the size of the customer, e.g., annual use, and assumes good information is available in the billing to support the analysis. The model is used to help choose the sample size *n*, to assess the expected statistical precision of any sample design, and to help formulate a sample design that is efficiently stratified for ratio estimation using case weights.

The model is used as a *guide* to the sample design, but the results of the study itself are *not* strongly dependent on the accuracy of the model.[5] Once the sample design is selected, the subsequent analysis of the data is based only on the sample design and not on the model used to develop the sample design. The resulting estimates will be essentially unbiased in repeated sampling and the confidence intervals will also be valid, provided that the sample design has been followed to select the sample customers. The results will be consistent with traditional sampling

---

[1] *Sampling Techniques*, by W. G. Cochran, 3rd. Ed. Wiley, 1977.

[2] *Model Assisted Survey Sampling*, by Carl Erik Sarndal, Bengt Swensson and Jan Wretman, Springer-Verlag, 1992.

[3] Wright, R. L. (1983), "Finite population sampling with multivariate auxiliary information," *Journal of the American Statistical Association*, **78**, 879-884.

[4] The report can be downloaded from the webaccount http://www.calmac.org/calmac-filings.asp

[5] Other methods, called model-dependent sampling, are much more dependent on the accuracy of the model. Such methods are not commonly used in load research applications since they would be more difficult to defend than MBSS and conventional methods.

theory as found in texts such as Cochran's *Sampling Techniques* and consistent with standard load and market research practice.

2.3 Stratified Ratio Estimation

We assume that a load research study is to be conducted of a given population of N accounts in a given rate class or market segment. In the study, a load research interval recorder will be used to measure the use of electricity of each sample customer on an hourly basis. We let y denote any customer characteristic to be determined from the customer's interval load data, and we let x denote any suitable characteristic of the customer's usage that is known from billing data such as annual use

We define the population ratio *B* by the equation

$$B = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i}.$$

Here the summations are over the entire *N* units (e.g., customers) in the target population. We note that the population mean or total of *y* is equal to *B* times the population mean or total of *x.* The latter is assumed to be known from the billing data.

We assume that a sample of n customers is selected following a stratified sample design. For each sample customer we define the case weight w to be equal to the number of customers in the target population within the stratum containing the given customer divided by the number of customers in the sample within the given stratum. The case weight is used to avoid any bias that might otherwise arise from the different sampling fractions used from one stratum to another.

Using the case weight, we define the combined ratio estimator of *B* by the equation: [6]

$$b = \frac{\sum_{i=1}^{n} w_i\, y_i}{\sum_{i=1}^{n} w_i\, x_i}$$

Then, if desired, the population mean or total of *y* can be estimated as *b* times the population mean or total of *x,* known from the billing data.

Using the case weights, we calculate the relative precision at the 90% level of confidence in three steps:

1. Calculate the sample residual $e_i = y_i - bx_i$ for each unit in the sample.

[6] This equation gives the same result as the conventional stratum-weighted equation: $b = \frac{\sum_{h=1}^{L} N_h\, \bar{y}_h}{\sum_{h=1}^{L} N_h\, \bar{x}_h}$.

2. Calculate[7] $se(b) = \frac{\sqrt{\sum_{i=1}^{n} w_i (w_i - 1) e_i^2}}{\sum_{i=1}^{n} w_i x_i}$.

3. Calculate $rp = \frac{1.645\ se(b)}{b}$.

A 90% confidence interval for *B* is calculated using the equation $b \pm rp\ b$. A confidence interval for the mean or total can be calculated in a similar way.

A key advantage of the MBSS methodology is the ease of domains estimation. A domain is any identifiable subset of the population, e.g., the sites in a particular region or having a particular appliance or end-use. Domain estimation is the process of obtaining the results of interest for one or more domains. With the MBSS methodology, domains estimation is very straightforward. We usually calculate the case weights for each sample site to reflect the sample design and current population and then regard them as fixed for any domains analysis. Then we simply evaluate the preceding equations for the sample sites that are included in each domain.[8]

2.4 Summary

Extensive experience indicates that stratified ratio estimation is very effective in almost all load research applications. MBSS methods are generally grounded on the same principles as conventional sampling methods such as Dalenius Hodges stratification and Neyman allocation, but MBSS methods are specifically tailored to ratio estimation. Some methods for calculating sample sizes that load researchers have commonly used in the past can provide badly misleading results for ratio estimation. The MBSS approach addresses these problems and provides a coherent, consistent approach to both sample design and analysis. The MBSS methodology follows the life cycle of load research studies very nicely.

A bonus of MBSS methodology is its strength for multiple y variables and domains estimation. At the sample design stage, MBSS provides straightforward methods for assessing the statistical precision expected for various y variables and domains of interest from the associated error ratios. At the analysis stage, MBSS again provides straightforward methods for developing estimates and their statistical precision for various y variables and domains, and for estimating the associated error ratios. In the past it has been thought to be risky to report results for domains that were not factored into the sample design. MBSS methodology has shown that meaningful results can generally be developed for questions that arise later in the study, much after the planning stage.

---

[7] The conventional equation is $se(b) = \frac{1}{\sum_{h=1}^{L} N_h \bar{x}_h} \sqrt{\sum_{h=1}^{L} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2(e)}{n_h}}$ where $s_h^2(e) = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (e_i - \bar{e})^2$. Our equation assumes that $\frac{1}{n_h - 1} \sum_{i=1}^{n_h} (e_i - \bar{e})^2$ is approximately equal to $\frac{1}{n_h} \sum_{i=1}^{n_h} (e_i)^2$ in each stratum.

[8] In this analyses, a domain is any rate schedule.