USAID FROM THE AMERICAN PEOPLE | Regulatory and Energy Assistance

LOAD RESEARCH GUIDELINES[1]

Prepared by

Bosnia and Herzegovina (BiH)

Regulatory and Energy Assistance Project (REAP) Staff

USAID Contract No. EPP-I-00-03-00004-00, Task Order 5

Implemented by Advanced Engineering Associates International, Inc.

March 27, 2013

[1] These guidelines are made possible by support from the American People sponsored by the United States Agency for International Development (USAID). The contents are the sole responsibility of the author/s and do not necessarily reflect the views of USAID or the United States Government.

Disclaimer

The author's views expressed in this publication do not necessarily reflect the views of the United States Agency for International Development or the United States Government.

The participants in the Bosnia and Herzegovina Market Working Group Load Research Subgroup acknowledge that these Load Research Guidelines are based on the Load Research Manual, Association of Edison Illuminating Companies, United States of America, 2001.

Contents:

Sample Design and Selection……4
Sample implementation……....17
Data Processing……28
Data Analysis……35
Data Transfer……61

Sample Design and Selection

INTRODUCTION

The electric load in electricity distribution varies with time and place, and the power production and distribution system must respond to the customers' load demand at any time. Therefore, modern electricity utilities need accurate load data for pricing and tariff planning, distribution network planning and operation, power production planning, load management, customer service and billing, and finally for providing informational to customers and public authorities.

The load information most needed is how a customer or group of customer uses electric energy at different hours of the day, different days of the week and seasons of the year, what their share of the utility's total load is and how loads of different customers are aggregated in different locations of a distribution network.

Usually, the most frequent and certain measurements from customers are energy consumption from the billing meters. Consumers at the higher voltage level mainly have interval meters. However, entire population of customers at low voltage level will not have direct hourly readings for many years. The customer billing databases usually include: consumer's name – code, address – region (urban, sub-urban, rural), type of meter (one/three phase; single/two period meter), contracted power, monthly/annual energy and the pricing information.

Load research programs must obtain reliable data to determine the energy use and demand characteristics of specific groups of customers, called "target populations." Data for an entire population is often too expensive to obtain and cannot be interpreted in a timely fashion. However, with reasonable expense and timeliness, data can be collected and interpreted as a sample that is representative of the targeted population.

1. SAMPLING AND CLASSIFICATION

Because of the large number of customers, sampling is the only possible way to collect data and estimate the consumption characteristics. The correspondence of the utility's customer classification in the categories to samples for load research depends on how well the utility people understand the background of each sample.

1.1. Advantages of Sampling

Sampling techniques are employed in load research for several reasons: reduced cost, greater speed, greater scope and greater data quality. Each of these is discussed below.

Reduced Cost - The cost of collecting information from a part of population (sample) is much lower than the cost of collecting information from the entire population.

Greater Scope - Another aspect of reduced cost is the ability to cover greater scope with limited resources by sampling.

Greater Speed - Data collection and analysis can be performed more quickly with a sample than with a population. Even if load data were available for an entire population, most

systems could not process the information within a reasonable time, except for very small populations.

Greater Data Quality - The smaller volume of information obtained from a sample allows increased attention and analysis time for each element, helping to ensure the quality of the data collected and analyzed.

1.2. Sample Design Criteria

At the outset of a sample design, several criteria must be considered. Important criteria include the following:

Objectives
The load research to be developed depends a lot on the defined objectives (but also on the available data). Program objectives must be clearly understood before sample design and selection. They should be clearly defined by all participating departments. Once objectives are defined, the target population can be identified and the analytic approach and sample design developed.

Usually the main objectives for basic load research are consumers' characterization and load profiling. However, natural continuation of the load research program also includes the allocation and estimation of losses and the load factor estimation that could be defined as additional objectives.

Considering all objectives at the beginning of a load study may prevent the high expense of a second complementary sample.

Population Definition
In load research, as for any sample design, the target population needs to be defined. The target population is the full set of individuals to be characterized, for which estimates will be derived from the sample. That is, the target population is the group about which inferences will be made from the sample data. The entire population of LV consumers will not be covered by remote meters for many years; however, information about the load diagram for this population is very important for utilities as well as for system operators, regulators, market participants and consumers.

Frame Specification
The population to be studied is divided into non-overlapping sampling units. This list of sampling units from which the sample will be chosen is called the frame of the study or the sampling frame. Divergence between the frame and the target population creates the potential for sampling bias.

Once the sample frame is established, sample design and selection may begin. Sample design defines the type of sampling technique employed, the stratification procedure, and the sample size required.

Sampling Bias Considerations
Bias is the difference between the mean value of the estimate (over all possible samples) and the true value being estimated. Sampling bias is the bias that results from sampling procedures. The most clear-cut sampling bias occurs when some members of the population about which inferences are to be made are accidentally or purposely excluded from the frame, thus prohibiting them from being selected for the sample.

Consumers division

Utilities usually face the following dilemma: consumers can either be divided into a predefined group or to cluster them in accordance to their load diagram, if exists. In the first alternative a variety of diagrams may appear in each group, so the average diagram would not be truly representative of the group. In the second alternative, a prototype of the groups would be good, but in certain circumstances the customer's load curve classifies him as he/she belongs to the one class of customers (small commercial as example) but the same customer is identified as he/she belonged to other class of customer (the household as example) in the database.

Therefore, for utilities without previous load research it is easier and quicker to get useful results to start load research with *a priori* defined groups. One possible way to group LV consumers is: Domestic, Administrative institution, Banks, Industrial, Commercial, Hotels and restaurants, Public-lighting (including also monuments (and other public) night lighting), Education institutions, Others.

However, utilities could also consider dividing some groups into subgroups (e.g. Domestic with central heating, Domestic without central heating) as well as dividing by type of meter (single period or two periods), type of region (Urban, Suburban, Rural) and geographical position (North, Central, South).

The correspondence of the utility's customer classification in the groups for the load research depends on how well the utility people understand the background of each sample.

The practical criteria for customer grouping are according to experience:

1. The variance in one group of customers should be as small as possible;
2. The number of groups should not be too large;
3. The group should be representative; and
4. The group should be easily linked with the utility's database.

Network selection

Having in mind the possible multi-objective purpose of the LR study, different selection network criteria should be taken into consideration. The selected network should satisfy some of the following possible criteria: to be representative of the global distribution of consumption; to be uni-zone (only urban, only suburban, etc.); to have a large number of consumers in order to facilitate the choice of samples; and to have only one type of line in an observed area (only overhead lines or only underground cables). If the LR study does not include a multi-objective purpose of the research, the network selection criterion could be simplified (e.g., if the purpose of LR study is not a loss study, the criteria ''to have only one type of line in an observed area'' is not important).

Desired Accuracy

The desired accuracy should be determined for the study. The desired accuracy for a study depends on the objectives of the study and how the results will be used. For a given target population, a larger sample size is necessary to achieve a higher accuracy level for comparable sample design and estimation techniques. If only a general load shape is needed for a small rate class, lower accuracy may be acceptable. One has to establish a maximum error allowed (e.g. 5 %, 10%, 15% or 20%) and confidence interval (e.g. 90% or 95 %).

Data Requirements
Load study plans should include provisions to collect the types of data that satisfy the program objective – for example, kilowatt-hours, kilowatts (5, 15, 30, or 60 minute intervals), temperature, and/or demographic data – and ensure that sufficient data is collected. Collecting excess data can be costly and often wastes computer, personnel, and equipment resources.
The main required data for the Load Research Study besides the data from billing files is at least hourly load sample from a measurement campaign. The smaller a period of measurement is, the larger the meter data-storage capacity is required.

Equipment Requirements
The type of equipment needed to collect load research data should be chosen carefully. Reliable recording equipment is necessary to ensure that reliable data is available for each sampling point. The frequency of data collection will depend on the meter data-storage capacity, datum interval length, data-retrieval method, and cost of manpower availability.

2. SAMPLING TECHNIQUES

There are varieties of the sampling techniques such as, *inter alia*, Simple Random Sampling, Stratified Random Sampling, Multi-Dimensional Sampling, Systematic Sampling, Cluster Sampling. The technique chosen depends on the load characteristics to be estimated, the information available about the population, the uniformity of the population, and other sampling and budget constraints. A good sample should be a good representative of a target population. This means that each sample point represents the attributes of a known number of population elements.

Simple Random Sampling and Stratified Random Sampling are often used load research sampling techniques.

2.1. Simple Random Sampling

Simple random sampling selects individual units throughout the population at random; each unit has an equal chance of being included in the sample, and no unit can be selected more than once. Since the sampling units are chosen randomly, any method of sorting the population prior to selection will result in a random sample.

Simple random sampling has several advantages: each unit of the population has the same probability of being selected. Simple random sampling is the easiest sampling technique to perform and the most flexible during analysis. In load research, simple random sampling is used mainly for populations with relatively few customers or for populations where individual units have similar usage characteristics.

2.2. Stratified Random Sampling

Stratification is the division of a population into mutually exclusive, non-overlapping groups or subpopulations, called strata. Stratification can increase the precision of sample estimates for the population and/or reduce the overall sample size required if the individual groups (strata) are more homogeneous than the overall population. Stratification may also be used to ensure that a minimum precision level or sample size is obtained for particular subgroups of interest.

In stratified random sampling, the population is first divided into subpopulations based on the stratification variable(s). The stratification variables must be available for every observation

in the population. In load research, therefore, these variables are usually obtained from information found in billing records.

If the overall population size is N and the number of strata is L, the strata are formed so that

$$N_t + N_z + ... + N_L = N, \quad (1)$$

where $N_{t,}, N_z, \ldots, N_L$ is the size of population in strata t, z, …, L.

A simple random sample is then selected within each stratum.

3. STRATIFICATION PROCEDURE

Stratification, dividing a population into subpopulations that are generally more homogeneous than the overall population, helps reduces the sample size needed to meet precision requirements. The stratification procedure involves two main steps in determining the number of strata required and the boundaries of the strata.

3.1. Determining the Number of Strata

The number of strata can often be determined for stratified random designs by calculating the coefficient of variation of the sampling distribution for test designs with different numbers of strata. The coefficient of variation is equal to the standard error of the estimate divided by the mean. The coefficient of variation declines as the number of strata increases, but at a diminishing rate. The point at which the coefficient of variation decreases very little as another stratum is added indicates the number of strata needed. The optimum number of strata is also indicated when the total number of sampling units required for a given precision level does not decrease significantly.

When the sample is to be stratified on a categorical variable, such as rate class or blocks within a class (e.g., 0-350 kWh, 351-800 kWh), the number of strata chosen is based on the number of subgroups on which data is needed and the practical number of strata that can be used.

Judgment and prior experience will help determine the number of strata for samples stratified on numerical variables, such as billed energy or demand.

For samples stratified either on categorical variables or numerical variables, restraints on the number of recorders available for a study may affect the number of strata that can be used.

Due to simplicity of the calculation, a large number of strata are lot recommended. No more than six strata are recommended for designs with conventional estimation techniques. For many load research studies, two, three or four strata are used.

3.2. Determining Stratum Boundaries

The boundaries of the strata can be determined by one of several methods. Some of the methods are classical statistics methods, but some of them are based on clustering and the application of artificial intelligence. Statistics methods are easier to understand and apply thn advanced methods, but are less accurate.

The most common techniques are based on the Dalenius-Hodges procedure, also known as the cumulative $\sqrt{f\,u}$ method or just cumulative $\sqrt{f}$ method. The procedure gives approximately optimal (i.e., variance-minimizing) set points for the Neyman allocation. The

procedure is one that is well explained in many statistics textbooks and one that is well suited for determining stratum boundaries for accounting populations. This procedure allows greater efficiency in the sampling process compared to setting boundaries using judgment only. Using judgment in setting up stratum boundaries is an acceptable method but may not be always the most efficient.

In the Dalenius-Hodges procedure, the population is divided into short intervals based upon the stratification variable. Each interval has frequency f and interval length u.

The quantity $\sqrt{f\,u}$ or $\sqrt{f}$ (in the case of applying the cumulative of the square root of the frequency method) is summed over all the intervals, and this cumulative $\sqrt{f\,u}$ or $\sqrt{f}$ is divided by the desired number of strata to give the optimum length of each stratum. This length is Test Interval (TI1). The first stratum boundary is the one that is the closest to the TI1. The second stratum boundary is 2 x TI1, and so on.

The $\sqrt{f\,u}$ procedure works best when the stratification variable is strongly correlated with the variable of interest, which is not always true in load research. Therefore, in most cases the cumulative $\sqrt{f\,u}$ procedure should only be a guide to the strata boundaries.

3.3. Determining Sample Size

Each load research study should be evaluated individually to determine the appropriate sample size. Drawing too small a sample compromises the reliability of the data, and drawing too large a sample produces needless information at great expense. The sampling method, estimation technique, and required accuracy all influence sample size requirements.

In general, when more emphasis is placed on precision, the sample size must be larger. The sample size appears in the denominator of the variance formula; thus, direct trade-offs exist between sample size and precision.

In many cases, either the sample size or accuracy level is set by a regulatory agency. With a given sample size, the analyst must choose the best sampling design to maximize the accuracy level. Conversely, with a given accuracy level, the analyst chooses the sampling design that minimizes the cost (by minimizing the sample size) necessary to meet the given accuracy level.

Equipment costs also influence sample size. A reduced per-unit recorder cost either reduces the total project cost or allows an increased sample size (within a given budget). The analyst should consider all types of costs associated with any load research project.

Availability of equipment may also affect sample size. If insufficient resources are allocated to a specific project, problems may result if the sample design calls for more equipment than is available. If the sample size is too large and only the first 100 sampled customers are installed (due to equipment availability), then an installation bias becomes a distinct possibility. That is, even though the full sample was selected randomly, the first 100 customers may tend to be different in terms of load characteristics from the remaining customers in the original sample. The result is a form of sampling bias, since the final achieved sample is only the first 100. It would be better to reduce the sample size and relax the variance expectations than to risk installation bias.

Precision Factor

In order to have the total samples size (n), the concept of statistical precision as in simple random sampling is applied. The formula is:

$$r = \left(\frac{\sigma}{\mu}\right) \cdot \frac{z}{\sqrt{n}} \tag{2}$$

where:

- r is precision (max allowed error),
- σ is the standard deviation of the total consumption (kWh),
- μ is the mean of the total consumption (kWh),
- z is the standard normal variable (i.e. z equals 1.960 for a 95% confidence interval – CI and 1.645 for a 90% CI),
- n is the sample size.

From the equation (2), the total number of samples (n) can be determined for a given precision factor. It can be seen that the smaller the precision factor is, the greater the samples required.

To complete the calculations in equation (2), the standard deviation and mean as shown in the following equations (3) and (4) respectively were determined.

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2} \tag{3}$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{4}$$

where:

- σ is the standard deviation of the total consumption (kWh),
- n is the sample size,
- μ is the mean of the total consumption (kWh),
- X_i is the current sample value of total consumption (kWh).

<u>Number of samples</u>

The total sample size (n) calculated from equation (2) is,

$$n = \left(\frac{\sigma}{\mu}\right)^2 \cdot \frac{z^2}{r^2}\,. \tag{5}$$

The total number of samples of each group of consumers can be determined for a given precision factor and a confidence interval. Usually, the precision factor is 0,05 (5%), 0,1 (10%), 0,15 (15%) and 0,2 (20%) and the confidence interval is 90% or 95%. It is also important to point out that different stratum inside one group of consumers can have different precision factors in accordance to its impact on total load. A stratum with a major impact on the load will be treated with a higher precision factor

If the σ/μ ratio is very large (typically >1), according to equation (5), it will produce very large sample sizes. Therefore, it would be good if the σ/μ ratio were less than 1.

To understand better the influence of the above parameters on the total number of samples, let us look some examples.

Example 1: Assume there are two groups of consumers (Group 1 and Group 2) with a total number of 400 consumers and 1200 consumers, respectively in the groups. If the mean of the total consumption is the same in each group and the standard deviation of the total consumption is also the same in each group, it means that the ratio σ/μ will be the same. Suppose, the precision factor r is 10%, and the confidence interval z is 90% are the same for both groups of consumers.
Applying the equation (5), the total number of samples (n) is calculated and shown in Table 1.

Table 1: The total number of samples for two groups of consumers with the same μ, σ, r and z

Group	Np	μ (kWh)	σ (kWh)	σ/μ	r	z	n
Group 1	400	3600	3000	0,833333	0,1	1,645	187,92 (188)
Group 2	1200	3600	3000	0,833333	0,1	1,645	187,92 (188)

It is possible to conclude that the total number of consumers (population), N_p, in a group do not have any impact on the total number of samples, n.

Example 2: We consider the same two groups of consumers with the same mean of the total consumption from Example 1. Now, we change the standard deviation of the total consumption to 1500 (kWh) for the Group 1 and to 360 for Group 2. The precision factor and the confidence interval stay the same as in the Example 1. The total number of samples (n) calculated according to the equation (5) is shown in Table 2.

Table 2: The total number of samples for two groups of consumers with the same μ, r and z and different σ

Group	Np	μ (kWh)	σ (kWh)	σ/μ	r	z	n
Group 1	400	3600	1500	0,416667	0,1	1,645	46,98 (47)
Group 2	1200	3600	360	0,1	0,1	1,645	2,71 (3)

We can conclude from the obtained results that the smaller the σ/μ ratio is, the smaller the total number of samples is.

Example 3: Now, we consider only consumers from Group 1 with μ=3600(kWh), σ=1500(kWh) and 90% confidence interval. Assume the maximal allowed error (precision factor) is changed from 15% to 5%. The results of calculation are shown in Table 3.

Table 3: The total number of samples for Group 1 changing the precision factor

Group	Np	μ (kWh)	σ (kWh)	σ/μ	r	z	n
Group 1	400	3600	1500	0,416667	0,15	1,645	20,88 (21)
Group 1	400	3600	1500	0,416667	0,1	1,645	46,98 (47)
Group 1	400	3600	1500	0,416667	0,05	1,645	187,92 (188)

It is easy to notice that higher precision factor requires more samples in the group.

Example 4: We consider again the consumers from the Group 1 with the same μ, σ and r as in Example 3, but we change confidence interval on 95%. The results of the calculation are shown in Table 4.

Table 4: The total number of samples for the Group 1 changing the confidence interval and the precision factor

Group	Np	μ (kWh)	σ (kWh)	σ/μ	r	z	n
Group 1	400	3600	1500	0,416667	0,15	1,960	29,64 (30)
Group 1	400	3600	1500	0,416667	0,1	1,960	66,69 (67)
Group 1	400	3600	1500	0,416667	0,05	1,960	266,78 (267)

We can conclude that increasing the confidence interval and the precision factor, the total number of samples is drastically increasing.

It is also important to point out that sometimes the number of samples is defined on the basis of the money the company budgets to spend on sampling resources (meter cost and installation). For example, the company says "We have the money to buy 900 interval meters. Please, decide where to install them." Then, the exercise would be to decide a viable combination of the maximum allowed errors and confidence intervals for each strata in such a way that the total number of samples do not go above 900.

3.4. Allocation to Strata

After the number of strata, the strata boundaries and the sample size has been determined, the allocation of the sample units to the individual strata must be made.

There are two types of allocation techniques that are commonly used in load research: proportional and Neyman.

Proportional allocation assigns sampling points to each stratum based on the number of population units represented in the stratum. If stratum 1 represents 10% of the population, then 10% of the sample points are allocated to stratum 1. Proportional allocation closely approximates simple random sampling, except that it forces a more representative distribution of the sample by a stratification variable.

More frequently used in load research, Neyman allocation assigns sampling points to each stratum based upon the percentage of the total population standard deviation represented by the stratum. The Neyman allocation optimizes the allocation for a fixed sample size; i.e., maximizes the precision.

Neyman Allocation

The Neyman allocation is a sample allocation method that may be used with stratified samples. The purpose of the method is to maximize the survey precision, given a fixed sample size. With the Neyman allocation, the "best" sample size for stratum *h* would be:

$$n_h = n * (N_h * S_h) / [\Sigma (N_i * S_i)] \qquad (6)$$

where:

- n_h is the sample size for stratum *h*,
- n is total sample size (equation (5)),

- N_h is the population size for stratum *h*,
- S_h is the standard deviation of stratum *h*,
- N_i is the total population of the stratum *i*
- S_i is the standard deviation of stratum *i*.

The Neyman allocation is actually a special case of optimal allocation.

Example: The total sample size is a sample of 36 customers. The system has 200 consumers; half is in Stratum I, and half is in Stratum II. The mean and standard deviation of each stratum are shown in the Table 5. How many sampled customers should be in Stratum I, and how many should be in Stratum II?

Table 5: Mean and standard deviation of customers in the stratum.

Stratum	Mean values (kWh)	Standard deviation (kWh)
I	700	102,7
II	800	66,7

According to the equation (6), the sample size for stratum I is:

$$n_I = 36 * (100 * 102,7) / [(100 * 102,7) + (100 * 66,7)] = 21.825 ,$$

The total sample of 36 customers should consist of 22 customers from the Stratum I and (36 - 22) = 14 customers from the Stratum II.

3.5. Accounting for Data Loss

Sample size and allocation equations estimate the optimum number of sample points to provide accurate data on the classes studied. However, data will not be available for every customer during every time period in the study due to events such as equipment malfunction, human errors, customers moving, and customers requesting removal of recording equipment.

Many researchers account for data loss in the sample size equations by applying a data loss factor. The data loss factor increases the overall sample size by a fixed percentage. This percentage can be determined from the amount of data loss occurring during a prior study or from researcher judgment.

For stratified sample designs, the additional sampling points are then allocated to each stratum in the usual manner. This procedure assumes that data loss is a random occurrence, with equal probability of occurring in each stratum.

Another consideration in accounting for data loss is the sample size allocated to a specific stratum. If an individual stratum sample size is small, even minimal data loss may cause estimation problems. To avoid this occurrence, some researchers set a minimum number of sample points per stratum and increase the sample size accordingly after allocation. The minimum number is a function of the data loss factor and researcher judgment.

3.6. Post-Stratification

Most stratified designs in load research are based on historic billing data. During the study period, the billing characteristics of some sampling units will change. Once the data is

analyzed, these units may be better classified in strata other than those to which they were originally assigned. After some time, it will be possible to reclassify these units by a technique called post-stratification.

4. SAMPLE SELECTION PROCEDURES

When the sampling technique, number, and type of strata, sample size, and sample allocation have been determined, sample units are selected by the procedures described earlier in this chapter under Sampling Techniques. Two issues that must be carefully considered during sample design and selection are customer exclusions and alternate sample units.

4.1. Customer Exclusions

During sample design, the information desired is determined, and the target populations are specifically defined. The defined population may exclude certain classes of customers from the study. For example, apartments should be excluded when targeting owner-occupied housing. Other groups targeted for exclusion may be delinquent, inactive, seasonal, or low usage customers.

Unplanned customer exclusion is an exclusion made after the sample is selected and commonly occurs because of hardware incompatibility, meter access problems, or lack of customer cooperation. To avoid bias, a reasonable effort should be made to install every sampling point originally selected. All unplanned exclusions should be examined to determine if specific exclusion tendencies exist that might affect the reliability of the sample.

The percentage of customer exclusion is also a concern. If the percentage of unplanned exclusions is significant, bias may be minimized by adopting a sound method of selecting alternative sample customers. The implications of unplanned exclusions and methods to minimize them will be addressed later.

4.2. Selecting Alternate Customers

Once all efforts to install a particular metering site have failed, selecting sample replacements becomes necessary. Replacements are chosen in one of three ways.
The first method chooses the replacement at random from the same stratum.
The second method chooses a random primary starting value and a secondary (replacement) random starting value during initial systematic sample selection. In this procedure, two samples are selected at the same time. If the primary installation fails, substitute the assigned alternate into the sample.
The third method selects a replacement customer from a group of customers as similar as possible to the excluded customer.

5. SUMMARY and RECOMMENDATIONS

The steps in the sample design are summarized below and some recommendations related to the steps are given for the beginning of Load Research in the power utilities in Bosnia and Herzegovina.

1. Determine objectives

The main objectives for basic load research are consumers' characterization and load profiling.

2. Define population

Targeted population for the power utilities are the customers connected to the 0,4 kV network. In order to start Load Research as soon as possible, it is recommended that the power utilities apply prearranged groups of customers from billing files. This means the Frame Specification should be consisted of the following groups:

- Households
- Others at LV with no demand metering
- Public lighting.

3. Define subgroups

The defined groups in the step 2 should be further stratified into subgroups as follows:

- Households in:
  - urban area with district heating system
  - urban area with no district heating
  - suburban area
  - rural area

- Others at LV with no demand metering
  - Workshops
  - Shops and handicraft workshops
  - Catering (restaurants, small hotels, bars)
  - Service activities (hairdressers, beauty salons, car washes, ...)
  - Administrative buildings
  - Banks
  - Healthcare buildings
  - Schools, religion buildings, sport halls
  - Water supply stations
  - District heating pump stations
  - Other customers
- Public lighting.

Correct identification of households' customers with different type of heating is impossible without questionnaires. However, due to different rate of energy consumption, it is expected that costumers with the same/similar type of heating will be in the same strata.

If the consumption feature of the customers' subgroup is not influenced by geography, it is not necessary to classify customers according to their geographic location.

In the coming years, it will be good to adopt classification that will be acceptable also for Statistics Agency/NACE according to EU standards of classification of economy-related activities. However, in this moment the above proposed classification is recommended to those responsible for Load Research

According to further analyses it is recommended that information about type of meter would be available for the chosen customers in the sample.

4. **Define main characteristics of each subgroup from billing files**

   After the division of customers, the power utilities must determine the main characteristics of each subgroup from billing files (number of consumers in the group, mean annual consumption, variance/ standard deviation).

5. **Choose Sampling technique**

   Power utilities will use the Stratified Random Sampling so that each subgroup will be divided into the strata. The stratification variable will be energy. In case it is possible, an ancillary variable to better select strata could be contracted demand.

6. **Define the number of strata**

   Recommended number of strata is 4 in each subgroup.

7. **Define stratum boundaries**

   The stratum bounders will be determined by the Dalenius-Hodges procedure as it was described in 3.2.

8. **Define maximum allowed errors**

   For the first Load Research Study power utilities will apply:

   - The maximum allowed error of 10 % or better if the customer subgroup has significant impact on the total group's demand;
   - The maximum allowed error of 15 % or better if the customer subgroup has medium impact on the total group's demand;
   - Maximum allowed error of 20 % or better if the customer subgroup has minor impact on the total group's demand.

   The maximum allowed error could be different for each stratum in a subgroup.

9. **Define Confidence Interval**

   For the first Load Research Study power utilities will apply a confidence interval of 90% or better.

10. **Calculate the number of samples**

    The total number of samples (n) is calculated for each subgroup from equation (5).

11. **Allocate samples to strata**

    After the total number of samples for each subgroup has been calculated, the samples will be allocated to the 4 previously defined strata applying equation (6).

12. Select consumers

Finally, the power utilities will select customers to install meters according to the number of samples in each stratum. Statistical characteristics of chosen customers in the stratum should be close to statistical characteristics of all customers in the observed stratum.

Sample implementation

INTRODUCTION

The care exercised during sample design and selection should carry over into the implementation phase of any load research study. In general, sample implementation consists of three steps:

- gaining customer participants,
- installing the hardware, and
- collecting data.

Carefully performing each step helps ensure accurate overall study results. Subsidiary tasks, such as following a regular maintenance schedule for equipment, also help assure maximum data collection.

1. HOW CAN WE ACCESS TO CUSTOMERS?

Practical access to customers that will be involved in a measurement sample can vary from utility to utility and depends also on obligations prescribed by the regulatory body.
The participation of consumers in the samples for the load research according to the international practices could be organized:

- with Customer's approval for participation,
- without Customer's approval for participation.

The participation with Customer's approval

The practice of soliciting customer approval for participation in a load research sample varies widely from utility to utility. Some utilities ask customer to accept participation in the load research study that may be important to maintaining good customer relations and protecting the sample integrity. Customers who understand what to expect as participants are more likely to accept the study at the outset and more apt to continue as participants for the duration. In general, customers may experience minor inconveniences: a brief interruption of service when the metering device is installed, perhaps an additional meter-read visit by company personnel from period to period, and a request to complete a demographic survey. However, if the new meters have communication abilities, meter-read visit will no longer be necessary. Even, if the meter has no communication abilities, most of interval meter have memory enough to keep all the data recorded in the last 6, 12 or 24 months (sometimes, even more). Therefore, the customer will not be disturbed so frequently.

Some special studies, such as end-use metering, may require that hardware be installed inside the customer premises and monthly access to the equipment be allowed. Therefore, many utilities seek the customer's agreement to participate, even though, legally, such an agreement may not be necessary.

While encouraging customer participation takes time and effort, it usually results in customers who are more responsive and cooperative than if they feel they have no choice. This is important so that the integrity of the sample can be maintained throughout the duration of the study. Some utilities also offer nominal financial or merchandise incentives to induce test customers to participate in load research studies.

The participation without Customer's approval

Nowadays new technologies provide possibility of measuring and collecting data without direct contact with consumers. Therefore, it is not needed that the consumers know they are selected. Sometimes it is possible to conclude about load and applied electrical devices (e.g. electrical heating) based on electricity consumption without any contact with consumers.

Ideally, load research samples would be "blind"; that is, customers would not know they are in the sample. In general, a concern with any data-collection process is that the data collection itself can alter the behavior of the individuals being studied.

However, data-collection technology and access to customer premises can make it difficult to install metering samples without a customers' knowledge. In old houses the meter could be inside the house, so the utility personnel has to contact the costumer to read the meter.

There is also another issue concerns consumer data. Some countries regulation states that consumer data belongs to consumers and he should authorize its use. In this case, even if the meter is outside, the utility personnel should contact the user and ask his permission to use his data. In some other countries regulator has a mixed position that is: data belongs to consumers but the distribution system operator (DSP) and other agents may use the data for load profiling or other studies, providing that no individual records are divulged.

The solicitation process may be similar for residential, commercial, and industrial customers, although commercial/industrial accounts may require more time to install the required meters and/or collect demographic data.

In the practices, it is happened really very rare, but it is possible that some customers will refuse to participate after some time or to drop out from the sample for any reason. The corrective actions taken when a customer drops out of a study vary widely and often depend on the type and duration of the study being conducted. For instance, if the study is scheduled for one year and a customer drops out after six months, no action may be taken (in these cases, the original sample size may be increased to account for such occurrences). However, for long-term studies, it is important to have alternative solutions (consumers) that will provide the load characteristics as similar as possible to the load characteristics of the customer who left the sample. A consistent treatment for handling customer dropouts should be decided upon prior to sample implementation and should be based on the objectives of the specific program.

2. HOW CONSUMERS CAN BE CONTACTED FOR INVOLVEMENT IN THE SAMPLE?

If regulatory bodies require the study of the load research it is not needed that customers would be contacted for their involvement in the sample.

In the case of Bosnia and Herzegovina, the regulators required study of the Load Research, so special customer solicitation to participate in the sample is not needed. However, if utilities are planning to extended the study with additional objectives such as collecting demographic and survey data it is advisable to make customer solicitation or at least customer notification. Simply notifying the customer of their selection in the study, usually results in just few customer refusals.

The method by which customers can be contacted varies from study to study and from company to company. The most often ways of contacting potential sample customer are:

- contact by mail/email,
- contact by telephone or personal contact.

Mail contact without follow-up is inexpensive and easy, but can generate high rates of customer refusal. Much like any other piece of mail a customer receives requesting his participation; it can be easily lost or discarded. Since the level of internet usage in BiH is law it is hard to believe the email solicitation can work.

Contact sample customers by mail (letter), followed by telephone or personal contact, is much more successful. It is more apt to impress the importance of the study upon the customer and invites questions before a decision. Personal contact by phone or in person without using a letter is, of course, another option. This method usually expedites the customer-solicitation process and allows the customer to ask pertinent questions.

If metering device is installed outdoor it is not needed to contact the customer. However, using this method exposes the utility to potential complaints from customers, sample deterioration, and additional expense in selecting alternate customers and, subsequently, removing and reinstalling equipment. Also, collecting demographic information may be difficult or impossible with this method.

The rate of customer acceptance can hinge on the knowledge of the staff doing the solicitation. Marketing/customer service representatives or meter supervisors/technicians offer advantages at this stage. Marketing/customer service personnel are trained to deal with customers on a wide variety of issues. When the representatives are familiar with load research issues, they can be persuasive in gaining customer acceptance. Metering personnel can also perform test-metering inspection as they solicit customer participation. They can often install the test meters on the first call or identify metering problems that might preclude customer participation.

Much of the previous discussion also applies to end-use studies. However, by their nature, end-use studies involve other considerations:

- End-use samples are often not random, since the total population (e.g., all customers with central air conditioning) cannot be identified.
- These studies often involve demand-side management programs or rate design experiments.
- The customer's internal wiring may need to be changed to accommodate metering equipment. This action may increase utilities study costs and may increase customers reluctant to participate.

Since end-use studies are more costly, the customer solicitation and monitoring process is even more important.

3. COLLECTING DEMOGRAPHIC AND SURVEY DATA

The load research study can be extended by an objective such as collect demographic and other relevant survey data from sample customers. This information usually includes family demographics, housing characteristics, and number and type of appliances in residential studies. In industrial and commercial studies, heating and cooling requirements, floor space, employment, and electric end-use loads may be collected.

The collected data can be compared with data obtained in prior studies and with results from large-scale surveys for representativeness. These data allow the researcher to examine

differences in load shapes among subsets of sample customers with similar demographic and end-use energy profiles.

The survey can be conducted by mail (letter), by phone, in person, or by any combination of the three. Mail surveys are the least expensive but the response rate for the mail survey questionnaires can be low and the accuracy-of the responses can be suspect.

While personal and phone interviews are more expensive, they can produce more reliable and accurate information. Interviewers can probe, where necessary, and respond to questions. On-site interviews, in particular, give the interviewer the opportunity to verify customer responses by checking appliances and making measurements.

If in-house staff is unavailable to conduct interviews, an outside research firm can be retained to survey the customers. Contracting with a firm knowledgeable about the subject and experienced in the field can ensure more accurate and complete responses. Outside firms can usually complete the interviews faster than in-house staff and they free in-house staff for other duties. A disadvantage of using outside firms is their inability to answer customer questions about general utility operations.

Also, one of the possibilities is to organize this task in cooperation with university. University teacher could propose works to be developed by students, where part of them would be the inquires and another part the data analysis.

As it is mentioned above, there are many ways to organize collecting demographic and survey data and each utility will chose the most appropriate way for the case.

Examples of questionnaire for customers from different categories (residential, industrial and commercial) are given in appendix of this module.

4. METERING

As the customer-solicitation process is completed on an individual basis, the required metering/communications package can be installed. During the installation process, exclusions may occur among customers agreeing to participate. Exclusions should be avoided whenever it is possible.

4.1. Customer Exclusions

Exclusions can be caused by hardware incompatibility or an inaccessible meter location. If a meter cannot be installed because of its location, either a remote device can be installed away from the meter or the point must be excluded. Atypical installations add costs and time to sample implementation, but eliminating the point could skew the sample to users with specific load patterns and result in a sample that does not accurately reflect the population characteristics.

Lack of access to equipment because of indoor meters or dangerous pets may potentially exclude customer from the sample. Creative meter configurations and communication alternatives can circumvent these problems.

While excluding random sample points might be necessary, the load researcher should hesitate to exclude points without researching alternative installation procedures. Although these procedures usually add costs and delays, they yield results that are more representative

of the population and worth more to the utility. Some customer exclusions may be unavoidable because of time or cost constraints or customer relations.

4.2. Coding Metering Sites

Each sample point in the field must be easily identifiable to the administrative, meter, and load research departments. If each department identifies metering sites differently, administrative errors will occur. Therefore, during installation, each metering site is usually assigned a unique load research code number. The code number can be obtained from the billing system or assigned by the load research department. It may contain elements of other identification variables, such as service district, geographic division, customer class, project identification, or stratum assignment. The code should be supplied to all participating departments and should not change throughout the course of the study.

To avoid identification errors, the code must be included with the data each time the data are retrieved. Proper coding also facilitates equipment changes due to malfunctions, customer removals, and data interpretation during analysis. As more classes and strata are included in a load research program, a coding system that easily identifies a particular customer becomes increasingly important.

4.3. Meter Configuration

Meter configuration and communications equipment are chosen by the metering department based upon the following factors:

- Meter location,
- Meter type,
- Equipment compatibility,
- Data-retrieval method (i.e., remote versus on-site),
- Equipment availability,
- Ease of installation,
- Duration of load study.

Nowadays, the utilities use Advanced Meter (or Smart Meter). The system includes measurement devices and a communication network, public and/or private, that records customer consumption (and possibly other parameters) hourly or more frequently and that provides for daily or more frequent transmittal of measurements to a central collection point.

Since 1985, Smart Metering has evolved from Automated Meter Reading (AMR) to Automated Metering Infrastructure (AMI) that enables two-way communication between customers and utilities.

This new technology in metering infrastructure provides an easier way to implement the load research study. Metering equipment may have the capability to record load in defined interval period throughout the day. Common forms of interval data include 15-minute data and hourly data. In other cases, meter replacement may be necessary.

If access to the meter is limited, remote communications with a radio link or telephone line may be preferred over periodic on-site data retrieval. Utilities should consider the cost of data

collection along the project period and the eventual risk of data losing in the case of on-site data retrieval.

It is important that Load Research is involved with the selection and implementation of the utility's AMI system. This isn't always possible so it is imperative that Load Research professionals familiarize themselves with the utility's choice of AMI. Load Research groups should take pains to be involved in the early stages of the utility's AMI development and implementation in order to identify and mitigate potential problems and how they relate to interval data collection.

4.4. Will AMI Replace the Need for Load Research Sampling?

As advanced metering infrastructure (AMI) becomes prevalent, some have suggested that it will replace the need for Load Research sampling. The theory is that when interval metering is available on 100% of customers, analysis can be performed using data from everyone, so statistical analyses will not be needed.

However, there may be rare cases where this is feasible, for most cases complete data is not available and statistically representative samples will continue to be the appropriate analysis approach. The most obvious reason to continue to utilize samples is the simple fact that the AMI system will not cover the entire population of customers.

Another reason relates to the amount of data that will need to be manipulated if summation is used instead of samples. Even though Load Researchers pride themselves on the amount of data they are able to manipulate, handling interval data for million customers for an entire test year may simply be more demanding than the Load Research itself.

Sampling is also a cost effective way of estimating customer load characteristics without having to maintain costly data storage and retrieval infrastructure. In addition, AMI system deployments take years, so one must consider what you are going to do in the interim.

Load Research sampling techniques will continue to play an important role during the planning; deployment; and full operation of AMI and meter data management. Sampling techniques will also play an expanded role in the measurement and verification of demand-side programs that are enabled by AMI. Looking towards the future, sampling techniques will play an increasing role in developing demand-side products.

Another, more intriguing, reason to not throw out the statistics books just yet, relates to the end-use studies that will be enabled by AMI. Even if there is no longer a need to statistically sample customer meters, there will still be sampling opportunities around the customer's end-use within the home. As the home area network evolves and the end-use measurement devices come down in price, it is possible that Load Researchers will be able to determine the usage profile of various appliances and plug loads within the house. This data, overlaid with customer data, such as the number of energy efficiency measures applied in the home, can begin to provide some real insight into customer usage patterns.

Load research also makes sense in the case of all customers with interval meters because sometimes there are gaps in measure data cause by meter malfunction, communication errors, data recording failures, etc. In these cases, some market operations (like settlement) is based on load profiles drawn from Load Research studies.
Another use of Load Research is in the field of planning. For instance, suppose that the utility wants to install a substation (or secondary substation) to feed a given portfolio of consumers.

In this case, Load Research may also provide very useful information to estimate the expected load diagram, key information for a proper dimensioning of the substation.
Therefore, it is easy to conclude that there are numerous reasons for the Load Research even if theoretically AMI will cover whole customers' population.

5. COLLECTING DATA

Data Collection is an important aspect of any type of research study. Inaccurate data collection can impact the results of a study and ultimately lead to invalid results. Data must be collected with minimum loss.

Efficient data collection occurs when involved personnel understand their responsibility for retrieval, and when retrieval and maintenance schedules are established and adhered to.

Many departments usually are involved in the load research study. The meter department may install and maintain equipment and retrieve data, the computer department translates and processes data, and group/department for analyses selects samples and analyzes data. When load research responsibilities are more decentralized, the coordination of its various activities becomes increasingly important. Each department must understand not only their own responsibility but also how it relates to the responsibilities of other departments and the corporate goals of load research.

Way of Data Collection depends on the type of meters that could be:

- meters with on-site readings,
- meters with possibility of remote readings and
- meters with automatic communication abilities.

5.1. Frequency of Data Collection

In the case of the meters with automatic communication ability any frequency of data collection is allowed.

If for any reason power utilities use the meters with on-site readings and meters with possibility of remote readings (occasionally remote readings) the following approach should apply.

The frequency of data collection depends on the frequency of measurement and the meter data-storage capacity, datum interval length, data-retrieval method, and cost or manpower availability.

Data-collection frequency could be dictated by *the memory capacity* using for data storage. If the memory can retain 30 days worth of data, the data need to be collected at least every 30 days to ensure that data are not lost. However, the data collection should be scheduled with a safety margin to ensure that the memory capacity is never exceeded.

The interval length being recorded also affects the number of days of data that can be stored by the load research equipment memory. Smaller interval lengths, such as 5-minute increments, utilize more memory and decrease the number of days of storage capability. Thus, if analysis does not preclude longer intervals, such as 30 or 60 minutes, the days of storage can be increased. Data may be collected more frequently than memory limitations. Frequent data collection during the study period can avoid loss of larger amounts of data.

Data-collection frequency may also depend on the *schedule of retrieving the data*. The load research department may develop a schedule for remote access data retrieval to ensure a consistent workflow.

The method of data retrieval and the equipment reliability also influence the frequency of data collection. On-site retrieval by hand-held solid-state devices contributes to labor costs (perhaps overtime), so the desired frequency of retrieval should be weighed against its cost.

Availability of manpower for the task may also limit retrieval frequency. On the other hand, solid-state recorders with a telephone attachment can allow nearly any frequency of remote reading with a telephone attachment can allow nearly any frequency of remote reading.

The characteristics of using equipment, system measurement and data collection should be checked carefully before making a decision of frequency of data collection

5.2. Data Retrieval

Data retrieval could be manually or remote retrieved from meters.

If collection of data is manually, a reader copies data into the memory of a hand-held reader and later transfer data to the company's computer at the processing office.

Data in a solid-state recorder with remote retrieval do not require on-site retrieval unless a problem occurs with the communication or the meter.

With remote communications access to the customer premise (especially inside) is not necessary. Issues of customer refusal, sample point replacement, and equipment compatibility are eliminated if AMI is in place.

However, while AMI may work in an urban environment, the need for drive by readings may still exist in rural areas. Therefore, load research group has to account on possibility that some data collection should be done on-site.

5.3. Sample Rotation

Utilities may decide to take the measurement systems for collecting the data from sampled customers at the same place for many years or to move the measurement systems and to replace one or more sampled customers with customers that have a similar or different behavior in electricity consumption. Anyway, utilities should consider the cost of moving interval meters that is probably more expensive than installing new meters in other costumers. However, even if the measurement system stay at the same place it is not guaranty that the customer and its behavior will be the same. Customers moving in and out of a service territory cause sample attrition. Therefore, after several years of load research, the validity of sample must be checked periodically to ensure that selected customer is good representative of the population.

Changes in usage patterns, housing stock, and demographics also change the population. Over time, samples may deteriorate to a point where they no longer represent the population. In case of sample deterioration it is probably better just to install new meters in the new customers to be monitored.
Due to the importance of load research samples for load profiling, common approach is to check sample validity every 2 to 3 years

When doing the check it is important to check if the typical profiles are maintained from one year to the other, in order to check if there are changes in customer behavior.
To check the sample against the current population, a new sample is designed and compared to the sample in the field. In many cases, a complete change-out of the sample may not be required.
Selected customers in the sample may be rotated to reflect the current design. To minimize the cost and implementation delays, existing test customers may be reassigned to new strata. If applicable based on the new design, selected customers may be added or removed. In continuous load research studies, required by many regulatory commissions, sample rotation is necessary to keep the sample current with the changing utility population.

6. SUMMARY and RECOMMENDATIONS

Sample implementation, although appearing to be routine, must be carefully planned and executed in all load research studies. The accuracy of the study results rests partially on the ability to install as many primary samples as possible and to collect data on a schedule that prevents data loss. Periodically checking the validity of the sample is necessary to ensure accuracy of the load research study.

In the case of Bosnia and Herzegovina, regulatory bodies RSERC and FERC issued "General Condition of Supply" where it is clearly defined that the DSO has right to read, change and maintain meters. Regulatory Commission for Energy of Republic of Srpska (RCERS) and Regulatory Commission for Electricity in Federation BIH (FERK) have also obligated the DSO to make load research. Therefore, it is not needed any special plan how to access to customers. Sample implementation as a step of the Load research study could be done without customer's approval for participation. However, if the distributions are planning to extend the objectives of the load research study (e.g. making questionnaires) it will be good to get customers approval for participation in the sample.

Information that would be collected from customers in different categories by questionnaires could help in identification of customer’s behavior. Therefore the distribution should seriously think about the survey organization. Number of customers who will participate in the survey is not needed to be the same as the number of samples. However, it would be good that each customer from the sample would be involved in the survey.

AMI is gaining strength as the new technological platform for utility metering and data collection. Many utilities have begun evaluating the various benefits of this technology and the impact it will have on their current business processes and IT infrastructure. From a Load Research perspective, implementation of AMI systems brings with it the idea that additional customer information will be retrieved and maintained. The remote communication meters would have been preferable.

The utilities in BH have been working on AMI system implementation. Different solutions for AMI system have been identified. In the General Condition of Supply and the Metering Rules and Distribution Codes, the separate chapter considers metering issue determining metering system, installation, meter reading, maintenance and distribution employee access to the metering point. However, from the load research study it is important that

- meter has possibility to measure at least hour load (however, due to future researches it is good that the meter can measure also 15-minutes or half of hours),

- communication with the meters would be as reliable as it is possible.

Each sample point in the field must be easily identifiable to the administrative, meter, and load research departments. It is strongly recommended that each sample point in the field has unique code number. If each department identifies metering sites differently, administrative errors will occur.

When the sample is defined, it is wise also to define a list of reserve customers to substitute the first choice.

Also, it is important that sample implementation would be as similar as it is possible in each distribution utility in Bosnia and Herzegovina due to possibility of exchange and compare results.

Depending on the type of communication data retrieval could be organize on different way and with different frequency of data collection. However, if it is not possible to collect data every day, the data need to be collected at least every 30 days to ensure that data are not lost for longer period.

After several years of load research, the validity of sample must be checked periodically to ensure that selected customer is good representative of the population. Due to the importance of load research samples for load profiling, the sample validity should be checked every 2 to 3 years. The checking will provide information if customer in the sample keeps the same characteristics of behavior and he is still good representative of the observed group.

Data Processing

INTRODUCTION

After individual customer's load profile data have been retrieved from field equipment, they undergo initial processing, validation, editing, and storage. Each of these operations is critical to the success of any load study and must be completed carefully, before the test data for individual customers can be subsequently grouped and extrapolated to the appropriate universe.

If test data for an individual sample customer are incomplete or contain errors that are not detected through the validation process, or if data in need of editing are edited improperly or stored carelessly so that some data are lost, when all of the sample customer results are grouped and extrapolated to class proportions, serious distortions may be introduced in the final test results. The occurred error in the final results can be considerable.

Data-processing procedures depend largely on the different factors such as:

- types of metering and data-retrieval systems,
- the choice of field hardware and data-retrieval systems,
- hardware, and software for data processing,
- the demand-profile recording interval.

Due to many different factors that can influence on data processing it is not possible to describe a procedure that will be use as a strict guidelines.

Load diagram data may be extracted from each customer premises via radio transmission, telephone lines, power line carrier, or portable electronic reading devices. The extracted data can be read directly into a central office computer. Many utilities prefer to use a dedicated computer (translator) as a central processor and data editor before storage on the mainframe or corporate network.

1. DATA RETRIEVAL

The data retrieval or "translation" process collects the recorded field data and combines it with the customer identifier data typically contained in a master file on the processing computer. The identifier information is crucial to the translation process. For each customer, the identifier indicates the source of the demand and time data, and provides required processing information that does not appear on the recording device itself. Attributing data to the wrong customer could invalidate the test point because each sample customer usually represents a set of consumers in a specific size range in the total class population structure. The importance of correctly identifying customer data cannot be overemphasized to the field forces.

Typical data files contain the following information:

- Test customer's name and address,
- Customer's billing account number,
- Company operating division,
- Special test customer identification number,
- Billing meter identification number,
- Billing meter kilowatt-hour constant,
- Billing meter start reading,
- Billing meter end reading.
- Start date and time of the data collection,
- End date and time of the data collection,

- Calculated kilowatt-hours consumed for the collection period,
- Any known electrical outage occurrences,
- Demand (kW) by interval for the collection period,
- Any special information or instructions

After the data are retrieved from the field, they are run through the computer. The time to read the data depends on the retrieval-communication method and can vary widely. The operator must initiate data retrieved from portable electronic devices. Most data-retrieval computers can be programmed to collect data equipped with remote communications during off-hours, perform initial data screening, and then alert the operator to exceptions.

Initial data screening includes:

- summing the usage by interval and comparing the result to the meter readings, and
- comparing the number of intervals collected versus the number of intervals expected.

Many data-retrieval systems also provide data-editing and data-validation capabilities. At a minimum, they are equipped to perform data screening and to signal the operator when data retrieval fails or data fails certain preprogrammed check routines.

Independent firms, such as meter service providers and meter data management agents may perform data processing.

2. DATA VALIDATION

Data-validation checks some identify errors in the metering and retrieval process that causes the data collected not to reflect actual usage. The validation report should contain customer information and summary-usage data. The report mainly compares the summation of kilowatt-hours and elapsed time, extracted from the field data, with the billing meter kilowatt hours, start and end dates, and clock times reported by the field technician. Energy and time - interval discrepancies are apparent, so that the analyst can later determine whether to accept, reject, or edit the data. Other validation criteria, such as excess peaks or number of zero-use intervals, can be set in some systems to further enhance the validation process.

Validation could be done due to internal as well as external users.

In the first stage of load research validation is made for internal users and analysts.

Validation reports or files are typically produced on an exception basis. The analyst intervention will be requested only if data fails validation. If the validation process for a particular test site indicates discrepancies, the analyst must decide to re-retrieve, reject, or flag the data for subsequent editing. These decisions are based on both predetermined editing guidelines and analyst judgment.

After initial processing, the unedited profile data may be flagged and stored until the load research analyst is ready to further inspect and edit the data. The timing requirements for data editing are based on the needs of the internal applications and periodic load research analysis.

At this pre-editing stage, data inspection can be performed in two ways.

An analyst can visually inspect the data-validation reports, together with the associated test customer daily load profiles and decide whether to accept, reject, or edit the data.

Graphical data visualization techniques may help in quickly reviewed for validation by the analyst. LV consumers real diagrams are quite irregular – sometimes, it is quite difficult to decide if it is feasible or not, unless it presents an unacceptable peak or other strange values. The analyst must be familiarized with these kinds of diagrams.

According to some experiences, if inspection of a data validation for a particular profile reading period indicates that the total kilowatt-hours registered on the billing meter or meters falls within ± 1-2% of the value obtained from the recorder, and the elapsed times check within two 15- or 30-minute intervals, the profile data can probably be accepted without detailed examination. These validation thresholds (± 1-2%) could be very strict at the

beginning of LR. Therefore, specification of these validation thresholds is usually given after a first preliminary analysis of the acquired data, just to assure that these thresholds will not discard a large amount of records.
Most utilities use a combination of automated validation and manual editing. Data problems are typically detected automatically, but then must be fixed manually. Inspecting and validating the profile data soon after its retrieval from the field allows the field forces to promptly correct any metering, communications, or recording problems.
When the data from field are collected, load profile data are no longer restricted to internal utility use. The collected data could be interesting to energy suppliers for settlement and scheduling coordinators for forecasting as well as to many other bodies, including regulatory commisison. Therefore it will be useful to establish rules for validating load data as well as rules for editing and estimating data. Before establishing validation procedures, load researchers must understand both internal and external requirements.
Validation standards include, but are not limited to, the checks shown in Table 1.

Table 1. Standard Validation Checks

Validation Check	Purpose
Meter identification and customer identification check	• The meter ID and customer ID were reported correctly • Data are being reported for the correct meter and customer
Pulse overflow check	• Correct scaling factor in meter • Properly sized transformer • No hardware problem
Sum check	• Correct channels between meter and recorder • No pulse relay problems • Correct PT and CT ratios • Correct meter constants
Spike check	• No data transmission errors • No spike resulting from a meter test.
High/Low usage check	• Correct meter constants • No energy diversion • Meters are not fast or slow • No dropped phases.

3. DATA EDITING

Editing can sometimes salvage data from incomplete records or correct obvious data errors. There are many different sources of load profile data errors in the recording and subsequent data-gathering operation - some human, some electrical. These sources include:

- Defective wiring between meter and recorder,
- Incorrect meter or recorder pulse constants,
- Errors in recording meter readings, dates, and data-retrieval times,
- Solid-state memory capacity exceeded,
- Failure to record and adjust for power outages if battery failure occurs,
- Improper merging of customer profile data from adjoining reading periods, causing either overlaps or gaps in data,
- Incorrect programming of the meter.

Editing can be performed at several stages during data processing: during translation, or after they reside on the final database.

Constant vigilance and training of field crews and data-processing personnel can go a long way toward eliminating or minimizing data losses and the need to perform substantial time-consuming editing. The load research analyst can spend many unproductive hours trying to discover the cause of bad data and selecting a repair method when careful data retrieval and documentation could have prevented the error. The load research analyst should resist a temptation to over-edit, or smooth, the profile data.

The natural diversity of electrical energy usage occasionally produces strange demand patterns, e.g., zeros, low values, and extreme spikes, as when large electrical loads come on line. Reference to a test customer's demographic information, usually gathered prior to recorder installation, can sometimes explain unusual demand patterns, as can knowledge of erratic customer behavior.

Records of equipment malfunction should be produced during the very first stage of the processing of field data. Such information is important to the analyst during the initial data acceptance or editing phases. Equipment malfunction reports should also be provided promptly to the meter department so that corrective action can be taken and further data loss prevented. To guide maintenance procedures and track manufacturer's performance, a separate file on meter/recorder malfunctions should be developed, ideally by the meter department.

Some of the occurred errors could be obvious errors very easy identified. These errors could be caused by incorrect entry of the meter constant, date and time. These relatively obvious errors in are easily corrected.

Sometimes validating the data reports and graphical data it could be noticed that several hours or days of data are missing or are not time-synchronized. The load research analyst may decide to reject the data for that customer or to edit based on substituted the missing data by data from similar day and hours or averaging the data from few days.

If some bad data are identified, analysts may attempt to estimate an individual customer's data by substituting demand profile values from similar days or hours for that same customer, if possible. The substitution process involves a great deal of judgment, and one can never be sure that the inserted data are truly representative of the missing values. This process may consist of a simple substitution of data from similar days, or may include a combination of substitution and proration so that a match with the metered energy is achieved. This process requires caution to prevent the artificial creation of a new peak Utilities should develop guidelines to indicate the type of customers who require estimation of missing data.

In all instances of profile editing, the data should be reviewed after editing has taken place and before final acceptance of the data into the post-edit database to determine that the adjustments have properly corrected the problem.

In the first cycle of load research project data editing is mainly done manually while during the time usually editing software is developed.

3.1. Evidence/Edit Trails

If some customer data are edited, it is considered good load research practice to maintain edit trails - accurate records of all modifications made to each set of data. These records provide important information on the validity of any customer data that have been edited. Data status flags or some comments should be added to each edited date indicated that some corrections have been done on original data. The good practice is to keep the original unedited data separate from the final edited version.

3.2. Editing Schedules

The timeliness of the editing and review process depends upon requirements of external constituents as well as internal schedules for data analysis. Many utilities find continuous review and editing of profile data the most efficient process. Others may wait until the entire test is over (provided that the test lasts no more than one year) to edit and analyze the data. Waiting for an entire year's worth of test data to be accumulated before any in-depth editing or analysis takes place has dangers attached to it. Some data problems that could have been identified and resolved during editing in the early stages of the test may go undetected, and may result in a customer's data loss for the entire test period.

4. DATA STORAGE

After the individual customers' data have been validated and accepted, rejected, or edited, they are stored in a database for further analysis. Rejected data can be stored separately from the final accepted data, or a field in the database can flag the rejected data and eliminate them from subsequent analyses.

4.1. Network Databases and Systems

Advances in computer technology provide the possibility to involve advanced data processing in load research. Internally, the corporate network links all databases and applications. This linkage makes data from many sources available to an analyst. Systems that typically retrieve, validate and edit data are different from systems that analyze data. In many utilities, different departments deal with different problems and different data. Sometimes, these kinds of studies require to access data from the different departments.

Due to the quantity of data accumulated by load data collection, efficient management of the large databases is required. Mainframes are used to house the data and make data accessible to the desktop. Advances in technology have allowed migration from mainframe applications and storage to network servers. On-line storage of data is inexpensive. However, it may have confidentiality risks. Many years of study data may be cost-effectively stored on-line. However, due to the large volume of data, accessing data may be time-consuming. Therefore, older data are archived to off-line storage to maintain a historical record and provide access, if needed.

4.2. Data Warehouses

Besides storing profile data, the load researcher must consider retaining any customer demographics and weather information gathered during the field tests. These data may be retained in a separate database or data warehouse.

A data warehouse is typically a relational database organized by customer account. It may be the "data of record" within a company used for both internal and external purposes. Users access one database rather than several, disparate databases to obtain data required for any given application. The original data records should also be stored, as well as the edit trails. A user of the data need not understand the complexity of the load data, but is assured of problem-free data for analysis.

5. SUMMARY and RECOMMENDATIONS

After the customer's data have been collected from field equipment, they undergo initial processing, validation, editing, and storage. The collected data could contain errors caused by different factors, human or electrical. Some of the occurred errors could be obvious errors very easy to identify and correct. However, some of the errors in the customer's files could not be easy recognized and could cause even unacceptable results at the end of the analyses. Due to different factors that can influence on data accuracy there is not exact procedure in this stage of load research, but some general recommendations could be followed.

1. It should provide easy connection data from field (measurement data) with the attributing data (data from billing files) for each customer.

2. Data processing, validation and editing could be done manually or by software. In the early stage of load research it is done manually. However, automatic data processing, validation and editing can save a lot of time. Therefore, it is necessary to identify and define as more as it is possible rules that will be later used in automatic data processing.

3. It is useful to do inspection of data by checking the recorded files and by graphical checking. In this way it is possible to identify missing data, very low and very high values of load.

4. The quick and easy inspection of a data validation for a particular profile reading period could be done comparing the registered kWh for billing and recorded load for the same period. For the future development of LR project, according to some experience, if inspection of a data validation for a particular profile reading period indicates that the total kilowatt-hours registered on the billing meter or meters falls within ± 1-2% of the value obtained from the recorder, and the elapsed times check within two 15- or 30-minute intervals, the profile data can probably be accepted without detailed examination. However, in the first stage of LR project implementation higher deviations are acceptable.

5. If some bad data is identified, the load research analyst may decide to reject or to edit the data.

6. Data editing is usually based on substitution process. The analysts may attempt to estimate an individual customer's data by substituting demand profile values from similar days and hours for those same/similar customers. This process may consist of a simple substitution of data from similar days, or may include a combination of substitution and proration so that a match with the metered energy is achieved.

7. Utilities should develop instruction to indicate the type of customers who require estimation of missing data.

8. Before final acceptance of the data, the edited data should be reviewed to determine that the adjustments have properly corrected the problem.

9. The edited data should be marked in the data base and contained information about made corrections.

10. The good practice is to keep the original unedited data separate from the final edited version.

11. Continuous review and editing of profile data is recommended as the most efficient process.
12. Due to the quantity of data accumulated by load data collection, efficient management of the large databases is required.
13. Mainframes are used to house the data and make data accessible to the desktop. However, copy of all data should be storage in the other computer.
14. On-line storage of data enables access to the data from different utilities' departments. However, accessing data may be time-consuming and this kind of storage can be risky. After, few years of load research, off-line storage to maintain historical records and provide access is recommended.
15. Due to customer demographics and weather information gathering in the future period is planned, the load analysts should bear in mind creation of a database for storage this information. The data warehouse is typically a relational database organized by customer account.

DATA ANALYSIS

INTRODUCTION

Data analysis as an important step in load research that enables an analyst to expand gained knowledge on a sample by means of metering, to all customers who belong to the same class and to the system by using different mathematical tools. Generally, mathematical tools applied in load research data analysis can be divided into statistical techniques and advanced artificial intelligence techniques. Statistical techniques have been used more often due to the fact that they are relatively easier to understand and have a simpler mathematical formulation. However, modern advanced techniques such as artificial neural networks, evolutionary algorithms and fuzzy logic enable better extraction of information from the collected measurement data but also need a knowledgeable staff who will work on data analysis. Regardless which technique is used, the aims of these techniques are to expand sample survey demand to total population parameters, to evaluate the accuracy of the demand estimates, and to identify relationships between the primary variables of interest and independent variables.

1. VARIABLES OF INTEREST

Load research examines how, when, and why utility customers use electric energy. Customer demand by hour, day, and date is the primary variable of interest used to study energy usage. Demand data collected from sample surveys can be extrapolated to target population parameters on the basis of number of the customers in the total population. The primary objective of most load research studies is estimating class demand.

As it was recommended in the previous chapters, demand data can be collected by recording instruments over specified time intervals of 15 or 30 minutes and stored in this form. Data can be analyzed in the intervals as stored or aggregated upward; i.e., 15-minute demand data can be analyzed as 60-minute demand data. Calendar-month energy consumption is the aggregation of hourly demand for all hours in the month.

The study objectives determine the appropriate demand interval and the time periods (hours, days, and dates) to be analyzed.

Variables related to and used in the expansion and analysis of the primary variable of interest are also important. These independent variables are quantities whose values can be determined for the population without reference to other variables. Energy consumption, total or per customer, for a rate class or an identifiable component of a rate class, is the most-often used independent variable. Others include the number and distribution of customers, customer demographics, appliance saturations, types of business, square footage, and weather. These variables are used to design samples, to expand sample survey demand to population totals, to establish domains within a target population, and to analyze and apply the survey results.

2. DATA PREPARATION

Getting the best possible results from data analysis assumes that the sample validation has been done and reasonably complete data is available from sample customers. Sample validation assumes comparison of sample average consumption with the universe (population) average consumption for each type or stratum. In practice, some data for individual customers, sample populations, and target populations are usually incomplete or unsuitable for analysis.

2.1. Handling Missing Data

The criteria for acceptance or rejection of customer demand data is normally defined as part of the data editing and storage functions of load research. Individual demand and/or complete customer records are coded as valid or invalid for each data-retrieval period. The coding identifies for the analyst which data is available for analysis.

A significant loss of data affects both the means and the variances of estimated demand in an analysis. This is particularly true for small sample sizes. The decision to include or exclude customers with incomplete or invalid data depends on the extent of data loss, the variable of interest, and analysis technique being used.

In general, if data for a customer is invalid for the time period of interest, the customer is excluded from the analysis. The only exception to this may occur if the analyst determined that excluding the customer will result in a larger error than including an estimate of the specific value. This is a rare occurrence and usually occurs when the customer exhibits extreme usage characteristics.

When the estimation variable represents an average of individual customer demand, either for successive hours or for the same hour on specific days, and partial data is available, the analyst must determine guidelines for the amount of missing data that will be tolerated. These guidelines are generally based upon the percentage of data available and can vary depending upon the number of components being averaged.

Many of the estimation procedures require data not only for the time period of interest but also for time periods encompassing it. One example is the ratio estimation technique using calendar month energy usage as the expansion variable. If data for a particular customer are available for the time period of interest, but are incomplete for the expansion period, the analyst must determine whether to include the customer. Often, if a significant amount of data is present, the expansion variable can be estimated based upon either billed energy consumption or by prorating the existing data.

To prorate the data, divide the existing data by the number of hours for which the data were collected. Then, multiply the per-hour rate by the number of hours in the month. For example, suppose you have a customer with 200 hours of valid load research data for July, which shows energy sales of 4000 kWh. The prorated value of energy sales is (4000 kWh/200 hours), or 20 kWh for each of July's 744 hours. Multiplying 20 kWh x 744 hours= 14,880 kWh for July.

The primary concern in handling missing data should be to determine guidelines that will produce the most reliable results. While estimating all invalid data points results in greater data availability, the accuracy of the estimate is often difficult to quantify and, thus, the overall accuracy of the sample estimate impossible to determine. Conversely, by not estimating the invalid data points, the resulting amount of available data may be too small to produce dependable results. Also, estimation techniques may not be reliable for all types of customers.

Data loss occurs in every study, usually at random. As such, if accounted for in the initial sample sizing, data loss does not significantly affect the study results. However, if a particular group of data is lost, the study can become biased. For example, if recording devices do not record data properly for small-usage customers, all of the study's data for these customers is lost. In this situation, the estimated usage per customer would overstate the true population value.

Most load research projects use stratified sample designs. The danger with stratification is that there may be too few sample customers with good data in each stratum to draw accurate conclusions. For example, due to data loss, there may be only one customer - or no customers - left in a stratum. Combining two neighboring strata may aid the situation. In this case, customers initially selected from each of the strata are combined to represent the populations of both strata.

2.2. Post-Stratification

Most load research samples are stratified on the basis of energy consumption. During the course of a survey, some customers commonly migrate from one stratum to another; that is, their consumption increases or decreases significantly such that they are no longer a member of their original stratum. Some degree of migration is expected and reflects the normal changes in the population. However, significant customer migration may cause the variance of the stratum and, hence, of the entire sample to increase significantly from the design level. Post-stratification of these customers normally improves the variances of the estimates. Relocating sample customers between existing strata does not necessarily impact stratum weights or boundaries, but is reflected in stratum means and error statistics.

On a broader scale, post-stratification may redefine or reestablish stratum boundaries and weights. Stratification of an existing non-stratified sample of customers, for example, not only improves the statistics of the estimates, but also provides additional insight on the behavior of customers within the newly established strata.

Redefining stratum boundaries and weights and assigning sample accounts to the redefined stratum may be an option if changes occur in the size, usage, or distribution of customers in the universe. Load surveys are implemented on the basis that these kinds of changes rarely occur. If characteristics of the consumers in sample or in the universe are always changing, the load survey results will be worthless.

While post-stratification is not normally anticipated in a sampling plan, these procedures may be necessary, particularly if the sample has been in place for a number of years or major changes have occurred in the total population. Population demand can be calculated and error limits specified for post-stratified samples with the same procedures applicable to the initially selected stratified sample; however, the results may be biased since the sampling units were not originally selected from the redefined stratum. Nevertheless, the results are generally reasonable and, in many cases, more representative of the population.

2.3. Energy Consumption

For billing purposes, customer meters are read monthly. Meter-reading schedules assign individual customers to billing cycles, generally numbered sequentially, which define their reading dates. The number of days normally billed ranges from 28 to 32, depending on the number of days in the month and when holidays and weekends occur.

Monthly billed sales, reported by utility, usually differ from the actual energy used in the calendar month.

Energy consumption reported from load research surveys is the summation of (15-min, 30-min or hourly) demand for the period of interest and is normally reported for calendar month periods, although demand can be aggregated to approximate billing cycle energy consumption.

For a matter of consistency checking analyst may also compute the energy consumption (from load research) corresponding to the billing periods – 28 to 32 days.

The disparity between the billed population energy consumption and the sample energy consumption is important because most estimation techniques use the total energy or the average consumption per customer of the target population to expand or adjust the demand of the sample to the total population. The total energy or the average consumption per customer of the target population is used to expand or adjust the demand of the sample customers to the total population. Accurate demand estimates depend on a close correspondence of energy-consumption measurement in the sample and target population.

Discrepancies between calendar month and billing cycle kWh can be minimized by adjusting billed energy to approximate calendar month usage, or by using the billed energy of the sample customers to correspond with the cycle billing of the utility.

One of the simplest methods to adjust billed energy to approximate calendar month energy consumption is to average billed consumption for the current and succeeding month. This technique recognizes that most billing cycles in a month include consumption from the prior month and that customer consumption in the current month is not billed until the succeeding month. Other adjustments to billed energy consumption can be made if deemed necessary. For example, the number of days in the billing period needs to be adjusted to the number of days in the calendar month.

A second option is to adjust billed monthly consumption on the basis of bills rendered in the billing cycle(s) read at the end of the month. If accounts are assigned to billing cycles on a relatively random basis, i.e., billing cycles are not geographically concentrated or customer consumption otherwise skewed, the consumption per customer in these billing units should more accurately represent the calendar month than all customers billed in the month.

Several methods are available for adjusting sample customer usage to correspond with billing records. One of the most common procedures uses billed energy for both the sample customers and the population as the expansion or independent variable and sample customer demand at the time of interest as the dependent variable. This technique accounts for the fact that a specific billing cycle may or may not include the demand period of interest, but is usually still highly correlated with the demand. When this technique is applied, the billing cycles of the sample customers should be distributed throughout the month and thus, adequately represent the population. Another method is to simulate billing cycle consumption from the demand of the sample customers and compute an equivalent consumption per customer for the sample (weighted as necessary) "as billed."

While no method exactly matches billed and sample energy to a common timeframe, these techniques can yield more appropriate estimations. As a test of reasonableness, the correspondence of average consumption per customer for the sample and the total population should be examined and reasons for large differences explored. If the differences appear still to be related to mismatched timeframes, some other method of adjusting the timeframes may be required. If the differences appear to be related to systematic differences between the sample and population, based on a comparison over a broad enough time that billing cycles could not have a large effect, a ratio or regression estimator might be preferable to mean per unit.

3. REPORTING RESULTS

Since load research provides information for diverse applications throughout the company, results must be reported in the format and with the detail required by the user. The uses of load research include load profiling, load forecasting, system planning, cost-of-service, rate design, market segmentation and analysis, retail market settlement, and the evaluation of demand-side options, such as load control, conservation, and cogeneration. The information needed for each application varies.

3.1. Load Research Report

Data analysis can be done for individual customers, for the total subclasses/classes, for the system in different time periods (week, month, season), according to the type of day (work day, Saturday, Sunday, special days) and different geographic areas - whatever is specified by the analyst.

Content of the Load Research Report basically includes:

- Data that could be easily identified and calculated from the collected data in the measuring campaign,
- Estimation of the classes demand and system demand at the level of the population,
- Calculations of different factors, and

- Graphics (typical load diagram).

On the basis of recorded load data, the following values per customers, subclass and class, and consumption area in the observed time period are identified:

- Maximum load P_{max} (and maximum energy consumption W_{max}) ,
- Minimum load P_{min} (and minimum energy consumption W_{min}),

and determined

- Ratio P_{min}/P_{max},
- Average load P_{sr},
- Energy consumption W.

A table containing the above-mentioned values for the defined customer types can be constructed. Table 1 represents an example of the information collected from the sample.

Table 1 – Basic information from the sample (Legend: ct – customer type, No.c.samp – number of customers in the sample, No.c.pop – number of customers in the population)

ct		No.c. sample	No. c. pop	P_{max} (kW)	P_{min} (kW)	P_{min}/P_{max}	P_{sr} (kW)	W (kWh)
Domestic - total								
Domestic - strata	1							
	2							
	3							
	4							
Commercial - total								
Commercial - strata	1							
	2							
	3							
	4							
Industrial - total								
Industrial - strata	1							
	2							
	3							
	4							
…								

After data is collected at the sample level, different techniques provide the opportunity to estimate class demand, peak demand and system demand at the level of the population. Some simple statistical techniques that provide a possibility for estimation will be described later in the chapter. The estimated demand and energy per sub-classes/classes of customer as well as per system should be involved in the report.

The main reason behind the calculation of different factors is to obtain a ''fast'' characterization of each diagram type and, in this way, have a simplified picture of the customer typical behavior. On the other hand, they contribute to make clear relations among diagram load factors and customer's contribution to those factors. The typical factors are

- *Load factor* (lf) is the ratio of P_{sr} and P_{max},
- *Coincident factor within the sub-class/class* (cf) is the ratio of max sub-class/class load and the sum of the non-coincident max load of customers within the sub-class/class,

- *Peak occupation factor* (pof) is the ratio of sub-class/class load in the moment of max system load and max sub-class/class load,
- *Contribution to the peak factor* (cpf) is the ratio of sub-class/class load in the moment of max system load and max system load.

The factors are calculated for different customer type and day type in a defined time period. Table 2 shows characteristic factors.

Table 2 - Characteristic load diagram factors (Legend: dt - day type: work day, Saturday, Sunday)

	dt	lf	cf	pof	cpf
Domestic	W				
	Sa				
	Su				
Commercial	W				
	Sa				
	Su				
Industrial	W				
	Sa				
	Su				
…					

Furthermore, the simultaneity factor of the system can be calculated for each month and day type. The coincidence (simultaneity) factor (CF) is the ratio of the max system load and the sum of the non-coincident max load of customers:

$$CF = \frac{P_{max}(system)}{\sum_{i=1}^{n^{o}consum} P_{i\,max}}. \qquad (1)$$

These factors provide a description of diagram characteristics. The typical values attained for each customer type in the sample are generalized for the customers' population.

Graphics are often more readily understood by the reader than equations and tabular data and give an added dimension to reports. Differences in load curves of customer classes, differences in seasonal load curves, the correlation of energy consumption and demand, the relationship of load factor and coincidence factor, and the effect of temperature on class energy consumption and demand can be more readily seen in an exhibit than in a table. Load diagrams represent the typical behavior of every class of customer, both in terms of shape as well as of magnitude.

As mentioned, many different type of load curves and the relationships between different values of interest can be drawn. On the basis of recorded data, a basic report of load research usually contains the following diagrams:

- Load diagram for each customer type at the sample level,
- Normalized load diagram for each customer type,
- Estimated load diagram for each customer type at the population level,
- System (network) load diagram,
- Load duration curves,

- Diagram kWh for each customer type,
- Diagram for each customer type normalized to global consumption – represents the consumer's type weight in the global consumption (total system diagram),
- Group kWh diagram – facilitates the evaluation of the total energy of each subclass/class of customer, to verify agreement with the energy balance.

Detailed load research reports can also include analyses of demand and energy consumption related to temperature, square feet of floor space, family income, or other related variables that mainly depends on utilities', customers' or/and regulators' needs.
If the load research group has computer programs (software packages) and load research data stored in a computer-compatible format, the basic load research report as well as the report expanded with additional information can be relatively prepared quickly. The additional information can include the billing records of sample survey customers, total system demand for each hour of the month or year, weather data, business codes, customer demographics, appliance saturations, and total system statistics (such as energy consumption and customers by rate class, system-wide appliance saturation data, number of customers and average consumption per customer by specified business classifications or geographic areas, and system output monthly and annually). Reporting capabilities are constrained by the database itself and by the flexibility of computer software.

3.2. Timing

Timing is an important element in reporting load research. Delay can be costly if tariff rates must be designed with outdated or inadequate data. Both the user and the analyst must understand the time constraints of a load research study and establish realistic deadlines. When demand data is readily available from an existing survey, essential reports can be compiled within a week or more. If a survey must be designed, meters installed, and data collected and analyzed, a realistic deadline might be much longer.
While new technologies in metering and accessing data have reduced processing time, and computer programs continue to expand the analyst's ability to respond to data requests more rapidly, proper planning and established priorities are still essential for ensuring timely reporting.

3.3. Distribution

Load research reports are mainly prepared for the utility, commissions, and the market agents. The different departments are interested in different information from load research. The reports also contain information and conclusions that are interesting for the media and the common citizen.

4. EVALUATING RESULTS

Sample results, whether class demand, appliance saturations, or customer demographics, are subject to error. Statistical inference uses sample data to estimate population parameters and to quantify the probable error in the estimate. The analyst evaluates the results of the analyses, identifies, if possible, the reasons for not achieving the desired precision, selects expansion techniques best-suited to the data available, and uses the results of these analyses to improve the results of future sample surveys.

4.1. Comparing Estimation Techniques

There is typically no one best estimation procedure for a particular data set and analytic question. Different estimation techniques may be used, provided the sample design is appropriately reflected in the expansions. Mean-per-unit is the simplest and most easily understood estimation method and has a history of acceptance by utility commissions.
On the other hand, difference, ratio, and regression methods have a number of advantages, and are recognized as appropriate methods in many jurisdictions. All these methods rely on the assumption that demand and energy (or other auxiliary variables) are correlated.
One advantage of the methods that utilize auxiliary variables is that the methods can directly adjust to changes in customer consumption over time. In addition, the techniques can often give reliable results with small samples. More generally, these procedures often result in better precision for the given sample than mean-per-unit.
Indeed, if prior information indicates that improved precision can be obtained using auxiliary variables and the sample has been designed to meet a target accuracy criterion utilizing the auxiliary information and a particular estimation approach, the target precision might not be attained using a different method, such as mean-per-unit.
With adequate sample sizes and no unusual distributions or data anomalies, all the techniques should provide relatively close estimates with acceptable confidence ranges. Estimating demand using more than one method is an appropriate way to ensure that data problems or poor analysis is not unduly affecting the results. The analyst must understand all the techniques to avoid improper specification of models.

4.2. Actual and Desired Precision

Desired precision levels are defined at the time of sample design. It might seem desirable that all demand estimates for all hours of the day and all days of the year should meet the defined criteria. However, sample precision varies by hour.
In the United States, a sample designed to meet the Public Utilities Regulatory Policies Act of 1978[1] named ''PURPA Standard'' or ''90/10 standard'' at the peak hour may have worse precision at other hours, particularly during periods of low customer demand. The 90/10 standard means that sampling procedures must be designed with a statistically expected accuracy of ±10% at the 90% confidence level for loads during peak hours. A utility is not required to show that the resulting load curves have this same accuracy, but it is required to point out any significant deviations from expected accuracy.
The achieved precision may also not meet the design goal, even at the peak hour. Unusually poor precision, particularly at the times of monthly or class peaks, may cause concern and should be investigated. Some reasons for poor precision may be

- A sample too small for the method of estimation,
- A variance in customer demand higher than that predicted,
- Poorly defined strata,
- Stratum migration,
- Data loss.

[1] The accuracy level for collected information is addressed in the Code of Federal Requirements CFR 290.403 (b) which stated that "for loads during peak hours, sampling procedures must be designed with a statistically expected accuracy of ±10% at the 90% confidence level. A utility is not required to show that the resulting load curves have this same accuracy but is required to point out any significant deviations from expected accuracy.

Precision not meeting the design criteria during low load hours is of less concern. Because the demand at that hour is low, the relative precision may appear large, but the absolute precision is generally small in comparison with high load hours.

4.3. Analysis of Sample Reliability

Since population demand are estimated from relatively small samples drawn from the population, a valid concern is how well the sample represents its universe. Actual population demand are unknown, precluding direct comparisons with estimated demand. The representativeness of a sample must, therefore, be judged on the basis of auxiliary variables that are available for both the sample and the total population and correlate well with the variable of interest, class demands. In these respects, energy consumption per customer is an acceptable proxy for demand.

If the mean and standard deviation of the sample compare closely with that of the total population, the sample can be considered reliable with respect to those specific parameters. Energy consumption of the sample should correspond closely to the target population consumption (per customer), not only annually but also for each month of the year. The frequency distribution of sample customers by month (weighted if stratified) should approximate that of the universe.

Another way to test the data is against prior demand estimates for the same class. Monthly peak hourly kW estimates can be compared across years for consistency, given known changes in related factors. Values that differ from those expected can often be explained by differences in the weather, time the peak occurred, or customer migration between rates. Differences that cannot be explained by such factors may indicate problems with the reliability of the sample.

Samples are continuously reviewed to insure ongoing representativeness with the target population group. If samples continuously fall outside the acceptable limits, they are supplemented with additional sample points, or replaced.

Resampling and supplementing an existing sample are seldom possible in the short run. More realistically, the analyst can use experience and additional information to improve the estimates of demand. Some options include trying alternative estimation procedures, restratifying and reweighting to improve the results, and infusing data from a parallel survey (although this option is not statistically verifiable).

5. STATISTICAL TERMINOLOGY

Applying statistics to load research requires an understanding of commonly used terminology and symbols.

Measures of Central Tendency

Several different measures of central tendency are defined below:

- The *mode* is the most frequently appearing value in the population or sample.
- To find the *median*, we arrange the observations in order from smallest to largest value. If there are an odd number of observations, the median is the middle value. If there is an even number of observations, the median is the average of the two middle values.
- The *mean* of a sample or a population is computed by adding all of the observations and dividing by the number of observations.

If the data is normally distributed, the mean, mode, and median are the same. The most often measure of central tendency used in the load research is the mean.

Sample Means and the Population Mean

Any sample mean is only an estimate of the true population mean. If we took a second sample, and then a third, we might find a different mean each time. Imagine that we took one hundred samples and calculated the mean value for each. We would have 100 mean values. Some might be the same as each other, but many would differ. Now imagine you could take every possible sample from a population and calculate their means. They form what is known as a *sampling distribution of sample means*.
The important things about the distribution to remember are:

- The sampling distribution is a theoretical construct. We do not actually have to collect multiple samples - one will do;
- The mean of any single sample is just one value of the many possible values in the sampling distribution of sample means;
- Sampling distributions are at the heart of a several of parametric statistical techniques.

Larger samples tend to have a mean that is closer to the population mean than smaller samples do.
The difference between the sample mean and the population mean is known as *the sampling error of the mean*. The sampling error does not refer to mistakes made during sampling or making calculations. It just refers to the fact that statistics from a sample are only estimates of the true statistics of a whole population.

Variance, Standard Deviation and Standard Deviation of the Sampling Distribution of Sample Means - Standard Error

Variance is a measure of dispersal of a population or sample variable about its mean.
The standard deviation of a sample is the square root of the sample variance.
The standard deviation of the sampling distribution of sample means tells you about how representative of the population your sample is likely to be. A very tight distribution indicates that the mean of any sample will be close to the population mean. A very wide distribution, on the other hand, indicates that any single sample mean has a higher chance of being far from the population mean. For these reasons, the standard deviation of the sampling distribution of sample means is given the far easier name: *Standard Error* (SE)
Standard error tells us the average distance between each possible sample mean and the population mean. Standard error is affected by sample in a number of ways:

- Sample size - Larger samples have a smaller standard error;
- Sample variation - Samples with larger standard deviation have a larger standard error;
- A wide spread means a large SE

The variance of hourly demand y_i for customers *i* is estimated from a simple random sample of size *n* as:

$$s_y^2 = \sum_{i=1}^{n} \frac{\left(y_i - \bar{y}\right)^2}{n-1} \qquad (2)$$

where $\bar{y}$ denotes the sample mean.
The standard deviation of the sample is the square root of the variance

$$s_y = \sqrt{\sum_{i=1}^{n} \frac{\left(y_i - \bar{y}\right)^2}{n-1}}. \qquad (3)$$

For a simple random sample, the standard error of the mean is the standard deviation of the sample divided by the square root of the sample size $\left(s_y / \sqrt{n}\right)$.

Covariance and Correlation

Covariance measures the degree to which two variables increase or decrease together. The covariance between two variables *x* and y, s_{xy} or cov(x,y), is equal to:

$$s_{xy} = \sum_{i=1}^{n} \frac{\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{n-1} \qquad (4)$$

The correlation coefficient *(r* or p) is the ratio of the covariance between two variables to the product of the two variables' standard deviations

$$r = \frac{s_{xy}}{s_x s_y}. \qquad (5)$$

The correlation coefficient ranges from -1 to +1. The values of correlation coefficient higher than 0,5 could be accepted. If r =1, it means that linear correlation exists between y and x. The correlation coefficient is also the square root of the *coefficient of determination* (r^2), which describes the proportion of the variance of *y* explained by the independent variable *x* in a simple linear regression of *y* on *x*. If, for example, the coefficient of determination $r^2 = 0{,}25$ it means that only ¼ data of x is correlated to the variable y.

Finite Population Correction

Standard variance formulas apply when the sample size *n* is selected from an infinite population, or when sampling with replacement is applied. When samples are selected without replacement from a finite population, the variability is reduced slightly. This reduction is accounted for by applying the *finite population correction factor* (fpc) to the variance and standard error formulas. The finite population correction factor is defined as 1 - *(n/N)* for the variance and the square root of this value for the standard error calculation, where *n* is the sample size and N is the population size.
The effect of the fpc is minimal on the variance calculation when the sample size is small in comparison with its universe. When the sample size is large compared with its universe (approximately 5 percent or more), the fpc noticeably reduces the variance and should be used.

Distribution

The distribution of a variable in a population is the frequency with which any particular group or range of values occurs for the variable in the population. For example, the distribution of monthly energy consumption in a rate group would be the proportion of customers in each usage range, or the cumulative proportion of customers sorted in ascending order by monthly energy consumption.

Another example of a distribution is the frequency of occurrence of estimated population means calculated from repeated random samples drawn from a target population. This type of distribution - the distribution of possible estimates from repeated sampling - is known as the sampling distribution for the sampling and estimation procedure.
The sampling distribution is the basis for statistical inference formulas, such as hypothesis tests and confidence intervals.
Population distributions are described as normal (a bell-shaped curve symmetrical about the mean), skewed (nonsymmetrical), or bimodal. Distributions of customers in many major rate groups are positively skewed; that is, the mean of the population falls to the right of the mode, with a tail to the right representing relatively few customers with high usage.
With a skewed population, the distribution of sample means for small samples is usually not symmetrical. Nevertheless, as the sample size increases, the sampling distribution tends toward normality. The F and chi-square distributions are examples of skewed distributions.
The Central Limit Theorem states that, for large samples, the distribution of sample means (individual averages of repeated samples of the same size) is approximately normal, regardless of the distribution of the population. Hence, it is possible to derive, for any confidence level, an interval estimate for the parameter of interest based on the Standard Normal Table.
The confidence bounds for an estimated parameter θ at confidence level *(1-α)* are given by:

$$\hat{\theta} + \mathrm{z}_{\alpha/2}\,\mathrm{SE}\left(\hat{\theta}\right), \tag{6}$$

where

$\mathrm{SE}\left(\hat{\theta}\right)$ is standard error of $\hat{\theta}$,

$\mathrm{z}_{\alpha/2}$ value obtained from the Standard Normal Table corresponding to the probability $\alpha/2$ in the upper tail.

For small sample sizes, the standard normal deviate $\mathrm{z}_{\alpha/2}$ obtained from the normal distribution tables is replaced by the corresponding value from the Student's *t* distribution. For an estimate of the mean from a simple random sample of size *n*, the degrees of freedom for the *t* distribution is *n-1*. As the sample size decreases, the number of standard error units for a specified confidence level increases. For larger sample sizes, the *t* distribution is approximately the same as the normal distribution. Typically, a sample size of 30 is sufficient to ignore the difference between the normal and *t* distributions.
Other distributions are used in load research for hypothesis testing and inferential analysis. Among these are the chi-square distribution to measure the divergence between actual and expected frequencies (Chi-Square Goodness-of-Fit Test), the F distribution to test for the significance of observed differences among three or more means and for the significance of regression models, and the log-normal distribution applicable to the analysis of skewed populations or samples.

Bias

If the mean value of a sampling distribution is not equal to the parameter it is supposed to estimate, the estimator is said to be biased. Bias is a systematic error, and in load research, is often introduced by sampling procedures, measurement errors, data loss, or exclusion tendencies. Bias is measured by the difference between the mean value of the sampling distribution and the true population value.

Confidence – Confidence Interval

The mean of a sample tells us something about the sample. In practice, however, we almost always want to say something about the mean of the population from which the sample was drawn. It would be nice if we could assume that the population mean was the same as the sample mean but we can't, since samples are a only part of the population. If we took two samples, you would probably get two different sample means. They can't both be the correct population mean as that value is fixed. The truth is that we cannot know the true value of the population mean without measuring every member of the population and that is often impossible.

The best we can say about a population mean based on a sample mean is that there is a high chance that it falls somewhere in a given range. This range is the *confidence interval* of the population mean. When we report confidence intervals, we must pick a *confidence level* - that is how confident we are that what we say is true.

Confidence is a statistical measure of the certainty that a sample statistic will fall within a pre-specified band width of the expected value of the population parameter being estimated. The parameter is measurable characteristics of the population (e.g., a mean or a standard deviation).

The confidence level specifies how confident we are that the true population mean falls into the confidence interval. The confidence level is given as a percentage and depends on the case. It is usually set in the range of 90% to 99%. As the confidence level grows, the confidence interval becomes wider. So the range into which we are 99% confident that the population mean will fall is larger than the range for the 95% level.

The confidence interval is the range that we believe the population parameter falls into. This interval will contain the true population parameter with the specified level of confidence. A confidence interval can be calculated for estimated averages, totals, variances, percentages, etc.

Note that the wider the confidence interval gets, the less useful the mean is, as it could just as easily be anywhere in a large interval. Confidence intervals are calculated from a sample of data. They are based on the size and variation of the sample.

The confidence interval around an average (under normality assumptions for the distribution of y_i) is expressed as:

$$\bar{y} \pm z_{\alpha/2} \frac{s_y}{\sqrt{n}} = \bar{y} \pm z_{\alpha/2} s_{\bar{y}} \qquad (7)$$

where

$\bar{y}$ is the mean of the sample,

s_y is the standard deviation of the sample,

$s_{\bar{y}}$ is the standard error of the mean and

$z_{\alpha/2}$ obtained from the Standard Normal Table, is the number of standard errors corresponding to the desired degree of confidence.

The value $z_{\alpha/2} s_{\bar{y}}$ is named the *margin of error*.

Table 3 - Confidence level and the corresponding $z_{\alpha/2}$

Confidence level	$Z_{\alpha/2}$
99%	2.58
95%	1.96
90%	1.65
80%	1.28
68,5%	1.00

Precision

Precision refers to how close estimates from different samples are to each other. Precision levels can be stated in absolute or relative terms. The absolute precision is the familiar multiplier $z_{\alpha/2}$ (or $t_{n-1,\alpha/2}$) corresponding to the specified confidence level multiplied by the standard error of the estimation variable. This quantity is expressed as:

$$\text{ApsolutePrecision} = z_{\alpha/2}\, s_{\bar{y}}. \tag{8}$$

The relative precision is the ratio of the absolute precision to the estimated value $\bar{y}$ and is expressed in percentage form as:

$$\text{Relative Precision} = z_{\alpha/2}\, s_{\bar{y}} / \bar{y}. \tag{9}$$

In general, the desired precision is stated in terms of the relative precision.

Weight

In stratified sampling, weights are used to combine the observations from different strata to estimate the parameters of interest for the whole population. If the weights are defined at the stratum level, the weight may be calculated as the ratio of the stratum to the total population, typically in terms of customer counts, expressed as a fraction or percentage. In this case, the population mean is estimated as the weighted sum of the sample stratum means, using these stratum weights. The sum of the weights over strata is one or one hundred percent.

Weights may also be defined at the level of each individual in the sample. With this approach, the weight may be the number of individuals in the population represented by the sample individual, which is also the inverse of the sample inclusion probability. In this case, the population mean is the weighted mean of the individual observations in the sample, using these weights. The sum of the weights is the population size. Equivalently, in terms of calculating population means, these weights may be scaled so that their sum is equal to the sample size. With more complicated sample designs or estimation procedures, weights may be defined in other ways.

6. HOW TO ANALYZE DATA FROM STRATIFIED SAMPLES?

In this section, we describe how to analyze survey data from stratified random samples.

There are few standard statistical estimation techniques that are usually used in load research to expand data to total system parameters and to estimate reliability of the results.

When it comes to analyzing data from stratified samples, there is good news and there is bad news.

First, the bad news: Different sampling methods use different formulas to estimate population parameters and to estimate standard errors.

Now, the good news: Once you know the correct formulas, you can readily estimate population parameters and standard errors. And once you have the standard error, the procedures for computing other things (e.g., margin of error, confidence interval, and region of acceptance) are largely the same for stratified samples as for simple random samples.

Four standard estimation techniques are described in this section: Mean-per unit, Ratio, Difference and Regression.

The following notation is helpful, when we talk about analyzing data from stratified samples:

- H - The number of strata in the population,
- N - The number of observations (customers) in the population,

- N_h - The number of observations in stratum *h* of the population,
- $\bar{y}$ - The sample estimate of the population mean,
- $\bar{y}_h$ - The mean of observations from stratum *h* of the sample,
- s_h - The sample estimate of the population standard deviation in stratum *h*,
- s_h^2 –The sample estimate of the population variance in stratum *h*,
- n - The number of observations in the sample,
- n_h - The number of observations in stratum *h* of the sample,
- SE - The standard error (an estimate the standard deviation of the sampling distribution).

Mean per Unit

The mean-per-unit technique calculates an estimated parameter of interest (i.e. demand) per customer and expands this estimate to the target population by multiplying by the number of customers in the population.

Measures of Central Tendency

Formula that can be used with stratified sampling to estimate a population mean is:

$$\bar{y}_{st} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h = \sum_{h=1}^{H} W_h \bar{y}_h \tag{10}$$

Note that $W_h=N_h/N$ is the sampling fraction (weight). Thus, to compute a sample estimate of the population mean we need to know the sampling fraction (i.e., we need to know the relative size of each stratum).

The Variability of the Estimate

The precision of a sample design is directly related to the variability of the estimate, which is measured by the standard deviation or standard error. Assuming that the sample method is stratified random sampling without replacement the standard error SE is

$$SE = \frac{1}{N}\sqrt{\sum_{h=1}^{H} \frac{N_h^2\left(1-\frac{n_h}{N_h}\right)s_h^2}{n_h}} \tag{11}$$

Confidence Interval

The confidence interval for the estimated parameter of interest per customer has upper and lower bounds given by

$$CI = \bar{y}_{st} \pm z_{\alpha/2} SE = \bar{y}_{st} \pm \text{margin of error} \tag{12}$$

Total

The total population parameter of interest $\hat{T}_y$ and its variance are estimated by the products of the number of customers and the sample statistics:

$$\hat{T}_y = N\,\bar{y}_{st} \tag{13}$$

$$s^2_{\hat{T}_y} = N^2\, s^2_{\bar{y}} \cdot \tag{14}$$

The following example is a didactic example, just to explain application of above equations.

Example: Let say the number of domestic customer is 200, and they are divided in two equal groups (strata) of the customers. In the measurement campaign 30 customers were selected, 15 in Stratum 1 and 15 in Stratum 2. Note: The number of customers is not chosen in accordance to the statistical calculation. The numbers is a mere example.

Table 4 – Sample measurements

Stratum 1 kWh	155	295	240	170	135	250	300	280	75	45	120	150	165	110	90
Stratum 2 kWh	305	400	380	325	310	420	390	320	330	405	365	335	315	305	360

Using sample data, estimate the mean reading achievement level in the population. Find the margin of error and the confidence interval. Assume a 95% confidence level.

Solution: Previously we described how to compute the confidence interval for a mean score. We follow that process below.

- Identify a sample statistic. For this problem, we use the overall sample mean to estimate the population mean. To compute the overall sample mean, we need to compute the sample means for each stratum. The stratum mean for the Stratum 1 is equal to:

$$\bar{y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}\bar{y}_i = \frac{\sum 155 + 295 + \cdots + 90}{15} = 172 \text{ kWh}$$

  The stratum mean for the Stratum 2 is computed similarly. It is equal to 351 kWh. The overall sample means is:

$$\bar{y}_{st} = \sum_{h=1}^{H}\frac{N_h}{N}\bar{y}_h = \frac{100}{200}172 + \frac{100}{200}351 = 261{,}5 \text{ kWh}$$

  Therefore, based on data from the sample strata, we estimate that the mean reading achievement level in the population is equal to 261,5 kWh.

- Select a confidence level. In this analysis, the confidence level is defined for us in the example. We are working with a 95% confidence level.

- Find the margin of error. To compute the margin of error when the sampling distribution is approximately normal, the key steps are shown below:

- Find the standard error of the sampling distribution. First, we estimate the variance (s_h^2) within each stratum. And then, we compute the standard error (SE). For Stratum 1, the within-stratum sample variance is equal to:

$$s_1^2 = \sum_{i=1}^{n_1} \frac{\left(y_i - \bar{y}_1\right)^2}{n_1 - 1} = \frac{(155-172)^2 + \cdots + (90-172)^2}{15-1} = 6777{,}86\,(\text{kWh})^2$$

The within-stratum sample variance for the Stratum 2 is computed similarly. It is equal to 1590 $(\text{kWh})^2$.

Using results from the above computations, we compute the standard error (SE):

$$SE = \frac{1}{N}\sqrt{\sum_{h=1}^{H} \frac{N_h^2\left(1-\frac{n_h}{N_h}\right)s_h^2}{n_h}} =$$

$$= \frac{1}{200}\sqrt{\frac{100^2\left(1-\frac{15}{100}\right)}{15}6777{,}86 + \frac{100^2\left(1-\frac{15}{100}\right)}{15}1590} = 10{,}89 \text{ kWh}$$

Thus, the standard error of the sampling distribution of the mean is 10,89 kWh.

- Find the critical value. The critical value is a factor defined according to the z score and Normal Distribution (Table 3). We find that the critical value is 1.96 for 95% confidence level.
- Compute margin of error (ME):

$$ME = 1{,}96 \cdot SE = 1{,}96 \cdot 10{,}89 = 21{,}34 \text{ kWh}$$

- Specify the confidence interval. The range of the confidence interval is defined by the *sample statistic ± margin of error*. And the uncertainty is denoted by the confidence level.

Therefore, the 95% confidence interval is from 240,16 to 282,84 kWh, and the margin of error is equal to 21,34 kWh. That is, we are 95% confident that the true population mean is in the range defined by (261,5 ± 21,34) kWh.

Ratio

Ratio estimation is a technique that can take advantage of the correlation of the variable of interest y with another variable x to obtain increased precision. When x and y are sufficiently correlated, the relative variance of the estimated ratio is less than the relative variance of either x or y alone. That is, knowledge of x provides information about y. As a result, the ratio estimator has better precision than the mean-per-unit estimator, which ignores the information about x.

Energy consumption for the sample and target populations must be for the same time period, usually the monthly billing period or the calendar month in which the demand of interest fall.

Formulas to estimate the ratio, mean, total, and variances in the case of the stratified random sampling follow. Detailed equations for these calculations could be found in Cochran, W. G. Sampling techniques, 3rd edition. John Wiley and Sons, New York from 1977..
From the sample data, we compute sample mean of the variable in stratified random sampling method as

$$\bar{y}_{st} = \sum_{h=1}^{H} W_h \, \bar{y}_h , \tag{15}$$

$$\bar{x}_{st} = \sum_{h=1}^{H} W_h \, \bar{x}_h , \tag{16}$$

where H is the number of strata, W_h is stratum weight. N is the number of units in population, N_h is the number of units in stratum h, $\bar{y}_h$ is the sample mean of variable of interest (demand) in stratum h and $\bar{x}_h$ is the sample mean of auxiliary variable (energy) in stratum h. The combined ratio estimate is

$$\bar{y}_{RC} = \frac{\bar{y}_{st}}{\bar{x}_{st}} \bar{X} = \hat{R}_c \, \bar{X} , \tag{17}$$

where $\bar{X}$ is the population mean of auxiliary variable.

The population total estimate of variable of interest is

$$\hat{T}_{y_{RC}} = \hat{R}_C \, X , \tag{18}$$

where X is the population total of auxiliary variable.

The variance of combined ratio estimate for the sample mean and the total is

$$V(\bar{y}_{RC}) = s^2_{\bar{y}_{RC}} = \sum_{h=1}^{H} W_h^2 \, \gamma_h \left(s^2_{yh} - 2Rs_{yxh} + R^2 s^2_{xh} \right) \tag{19}$$

$$V\left(\hat{T}_{y_{RC}}\right) = s^2_{\hat{T}_{y_{RC}}} = \sum_{h=1}^{H} N_h^2 \, \gamma_h \left(s^2_{yh} - 2Rs_{yxh} + R^2 s^2_{xh} \right) \tag{20}$$

where $\gamma_h = \frac{1-(n_h/N_h)}{n_h}$, $R = \bar{Y}/\bar{X}$ is the population ratio, s^2_{yh} is the population variance of variable of interest in stratum h, s^2_{xh} is the population variance of auxiliary variable in stratum h and s^2_{yxh} is the population covariance between auxiliary variable and variable of interest in stratum h.
When first degree approximation is used in obtaining the mean square error (MSE) of a ratio estimate, it is known that MSE is equal to the variance.

If the variable of interest y is demand and the auxiliary variable *x* is energy, in the combined ratio estimate technique, the ratio $\hat{R}_c$ is kW demand divided by kWh in the month or billing period. According to the definition, the load factor is the ratio of average demand and the maximum demand:

$$LF = \frac{kWh\,(\text{in monthly or billing period})}{hrs\,(\text{in monthly or billing period})} \cdot \frac{1}{kW} \tag{21}$$

It is easy to notice that the ratio is the inverse of the load factor, ignoring the number of hours. The number of hours is not included in the ratio estimation formula since it is assumed that this value is the same for both sample and total population and cancels out in the expansion step.

Difference

The difference technique may be used rather than the ratio technique when *x* and y have a linear relationship that does not pass through the origin. The generalized form of the equation for the estimated average demand of the population using the difference technique is:

$$\bar{y}_d = \bar{y} + \beta_1\left(\bar{X} - \bar{x}\right) \tag{22}$$

where $\bar{y}$ is sample mean of the demand per customer for the interval(s) of interest, $\bar{x}$ is energy consumption per customer for the sample (as billed or calendar month), $\bar{X}$ is population mean of the energy consumption per customer for the corresponding period, β_1 is a coefficient relating incremental demand per customer to incremental energy consumption per customer. The estimated total demand is

$$\hat{T}_{y_d} = N\,\bar{y}_d\,. \tag{23}$$

If $\beta_1 = 0$, the equation is equivalent to the mean-per-unit estimate.
If $\beta_1 = \bar{y}/\bar{x}$, the result is equivalent to the ratio estimation technique.
If β_1 is assumed to be the slope of the regression between y and x, i.e., $\beta_1 = \text{cov}(x, y)/s_x^2$, the estimated variance of the average is calculated as

$$V(\bar{y}_d) = s_{\bar{y}_d}^2 = \left(1 - \frac{n}{N}\right)\left(\frac{1}{n(n-2)}\right)\left\{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \frac{\left[\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})\right]^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right\} \tag{24}$$

$$s_{\bar{y}_d}^2 = \left(1 - \frac{n}{N}\right)\left(\frac{1}{n(n-2)}\right)\left[s_y^2 - \hat{\beta}_1\,\text{cov}(x, y)\right], \tag{25}$$

where $\text{cov}(x, y) = \dfrac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{n-1}$ and $\hat{\beta}_1$ is the estimate of β_1 obtained from the sample.

The estimated variance of the total is

$$s_{\hat{T}_{y_d}}^2 = N^2\,s_{\bar{y}_d}^2\,. \tag{26}$$

Two previous equations assume that the difference between the sample mean $\bar{x}$ and the population mean $\bar{X}$ is small compared to the standard deviation of x.

Regression

Regression models are most commonly used to describe a relationship and to predict future values or values for particular conditions. Regression methods are adaptable to analyses and extrapolation of a subpopulation drawn from a primary sample. These methods connect the dependent variable y (usually demand) and the independent variable(s) x (e.g., energy consumption per customer). The simplest models use single dependent and independent variables, assume a linear relationship between these variables, and assume that error terms (variance) are homoscedastic (vary randomly and uniformly about the regression line). Multiple linear regression models can be constructed, and weighted by the least square methods rather than the more conventional ordinary least square regressions can also be used.

Linear regression techniques estimate a linear relationship between a dependent variable such as kilowatt demand at the hour of interest and an independent variable such as kilowatt hours, peak demand, weather, or electricity price. Linear regression procedures calculate the line that most closely fits all the data points that represent the observations. This line is obtained by choosing coefficients to minimize the sum of squared errors. The simple linear regression model is:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{27}$$

where y_i is the dependent variable (e.g., demand), x_i is an independent variable (e.g., energy consumption), β_0 is the intercept, corresponding to the value of *y* when x=0, β_1 is the slope of the line and ε_i is the error.

Multiple linear regression model estimate the demand with more than one independent variable without the intercept is:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \varepsilon_i . \tag{28}$$

If the intercept term β_0 is included, the model is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \varepsilon_i, \tag{29}$$

or, polynomial model is:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \beta_3 x_{1i}^3 + \beta_4 x_{2i} + \beta_5 x_{1i} x_{2i} + \varepsilon_i . \tag{30}$$

Once the regression coefficients ($\beta_0, \beta_1, ...$) are estimated, the population averages of the independent variables are applied to the model to estimate the average demand. Thus, for the simple linear regression case, the estimated population means demand is

$$\bar{y}_{lr} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1, \tag{31}$$

where $\bar{X}_1$ is the population mean is energy consumption per customer, $\hat{\beta}_0$ is the estimate of β_0 obtained from regression, $\hat{\beta}_1$ is estimate of β_1 obtained from regression.

Multiplying $\bar{y}_{lr}$ by the total population N determines the total demand of the population:

$$\hat{T}_{y_{lr}} = N \bar{y}_{lr} . \tag{32}$$

For simple linear regression models without an intercept, the total demand for the population can be calculated directly as $\beta_1 X_1$, where X is the total energy consumption for the population in the month of interest. The variance of the mean is:

$$s^2_{\bar{y}_{lr}} = \left(1 - \frac{n}{N}\right) s^2_{\varepsilon} \left[\frac{1}{n} + \frac{\left(\bar{X} - \bar{x}\right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} \right] \tag{33}$$

where

$$s^2_{\varepsilon} = \frac{\sum_{i=1}^{n} \left[\left(y_i - \bar{y}\right) - \hat{\beta}_1 \left(x_i - \bar{x}\right) \right]^2}{n - 2}. \tag{34}$$

The variance of total is:

$$s^2_{\hat{T}_{ylr}} = N^2 s^2_{\bar{y}_{lr}}. \tag{35}$$

7. SPECIALIZED STATISTICAL APPLICATIONS

While the primary application of statistical theory in load research is to expand sample results to total population parameters, the analytical methods used for this purpose have specialized applications. Some of them are

- *Averaging Load Estimates Across Time*
- *Aggregating Class to Total System Demand*
- *Load Factor-Coincidence Factor Analysis*
- *Weather Normalization*

Averaging Load Estimates Across Time

For many types of analysis, demand averaged over hours or days are more appropriate than demand for specific hours on a specific day. While considerable time and effort are spent analyzing demand on system peak days for both cost allocation and forecasting purposes, average weekday demand by month or season are better adapted to forecasting models designed to predict demand for all hours of the year. Average demand over on-peak and off-peak hours are needed to evaluate the effect of time-of-day rates on customer usage patterns. Class load curves of the average of the five warmest days of the summer may be more typical of customer peak use than the single warmest day, since the effect of abnormal weather or customer behavior will be diluted.

Average demand, whether for the same day in successive weeks, all weekdays in the month, or successive hours in a day, can be estimated by mean-per-unit, ratio, or regression techniques. Assuming each of the sampling units has complete data over the period of the average, standard errors are calculated and have the same meaning as those for a single interval estimate. If some data is missing, care must be taken in how the averages for each sample point are constructed before calculating the sample average of these time averages. To avoid errors related to the timing of missing data for individual customers, it may be best to impute the missing data. In this case, calculating the standard errors of the sample average as if all the data had been complete originally will somewhat understate the true standard

errors. A rough adjustment is to inflate the calculated standard errors by the square root of the ratio of the full sample size n_{full} to the average number of points n_{avg} with nonmissing data:

$$SE_{true} \cong \sqrt{n_{full}/n_{avg}}\, SE_{full} \tag{36}$$

Even with non-missing data accounted for, average demand usually vary less than single interval demand. Thus, the variance of the averaged estimates will be relatively lower than those of single interval estimates. Consider that for a stratified sample of residential customers, the relative variance of demand at the time of system peak can be more than five times larger than the relative variance of the average annual customer consumption (the aggregated customer demand for 8,760 hours).
Alternatively, if the analysis of average weekday demand is based on five days of data, the standard error of the results will be reduced due to the increase in sampling points being utilized.

Aggregating Class to Total System Demand
Analyzing system peak demand by rate class is fundamental to developing cost-of-service allocation schedules. Whether allocations are based on the peak responsibility (class contributions to the annual system peak) or the average of twelve (contributions to the twelve monthly peaks) concept, the first step in the analysis is aggregating class demand and losses to the total system demand.
For many rate classes, contributions to peak are estimates subject to quantifiable error, and electrical losses must be estimated for each rate class based on voltage level, temperature, demand, and other factors. The resulting summation of the class loads and losses will usually fall short or exceed the recorded system demand.
While there is no standard procedure for adjusting component demand to total system demand, the analyst can use sampling theory and the confidence intervals associated with the contributions to peak of the major rate classes to make any necessary adjustments.
One option is to accept the initial analyses without adjustment. This approach may be practical, particularly if the difference between total load and the sum of the parts is small. Allocating the difference to rate classes would have only a minimal effect on the percentage distribution of the total load to each of the rates. (Note that the difference between the total load and aggregated load will generally be smaller when dealing with average value (average monthly peak) than with single value (monthly peak). This is supported by both sampling theory and experience.)
Proration is the most direct method to make class adjustments to the total system load. Some options adjust the class load research demand estimate by adding to it a proportion of the difference between the sum of the class load research and the known system demand. The variance is minimized in this case when the proportion used is equal to the ratio of the variance of the class load research estimate to the sum of all the variances of the class load research estimates.
Adjustments to peak day loads for the full day, or for the peak and adjacent hours, may also be valuable to the analyst.

Load Factor - Coincidence Factor Analysis
In addition to statistical techniques, such as regression and ratio analysis, and methods appropriate for estimating population parameters from sample surveys, the analyst should be aware of the unique statistics conventionally calculated from energy consumption and demand data and their uses.

Several of these statistics are described in this section.
Annual, monthly, and daily load factors are the ratio of average demand over the period and the peak demand recorded during that period. Load factors are calculated for individual customers, customer classes, and the total system. Load factors of individual appliances are also of interest.
Demand may be for 15-, 30-, or 60-minute intervals. The general formula is

$$LF(\%) = \frac{Total\ energy/number\ of\ hours}{Peak\ demand} 100\% \tag{37}$$

A statistic useful to forecasters, not technically a load factor, is the ratio of average energy consumption divided by demand at the time of system peak.
Coincidence (Simultaneity) factors relate the class peak demand for a specified period, e.g., month or year, to the class noncoincident peak demand (summation of the individual customer peak demand) during the period.

$$CF(\%) = \frac{Class\ peak\ demand}{Noncoincident\ peak\ demand} 100\% \tag{38}$$

Coincidence factors are also calculated as the ratio of system peak demand to the sum of the component class demand, regardless of time of occurrence. A coincidence factor cannot exceed 1 (or 100%).
Load factors and coincidence factors for individual customers have significant considerations in rate design, particularly for large industrial customers. Class load factors have applications in rate design, forecasting, and system planning. The relationship between the load factor and coincidence factor provides an estimate of the probable contribution to the peak of individual customers based on their monthly load factors. An illustration of the load factor-coincidence factor relationship (Bary Curve) is shown in Figure 1.
Other energy and demand relationships and ratios of interest to market analysts and for ratemaking are monthly load factors by energy usage level and type of business.

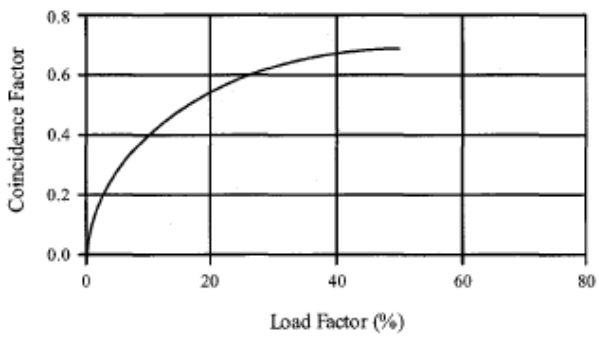


Figure 1 - Load Factor-Coincidence Factor Relationship

Weather Normalization
Weather variables are independent variables that affect both sales and demand. Therefore, weather normalization, or more broadly the evaluation of the effect of weather on sales and demand, is necessary to assess the effect of abnormal weather on revenues. These tools can be used for estimating the total impact of heating and cooling loads on system outputs and demand that can provide input to decisions on the need for load control programs.
The principal tool used to analyze the effect of weather on system and class loads and sales is regression. The primary problems are in model specification and in defining the variables to be included in the calculations. Computers have currently a large capacity to collect and store data and to calculate a wide range of regression curves from the data provided. The analyst is

responsible for ensuring that the results are reasonable and useful as a forecasting tool. If the regression is used for normalizing sales and demand, the weather variables used to construct the regression equation must have counterparts for normal weather. Both linear and nonlinear models should be tested to ensure the best fit.

8. SUMMARY and RECOMMENDATIONS

The aims of data analysis in load research are to expand gained information on samples by metering to total population parameters in the defined customers' types or in the system, to evaluate the accuracy of the estimates, and to identify relationships between the variables of interest.

The main objective of the load research studies is estimating demand of different customers' type.

The primary variable of interest is customer demand by hour, day, and date. Two secondary variables, the most often used are energy consumption (total or per categories/customer) and weather.

Mathematical tools applied to data analysis are divided in statistical techniques and advanced techniques based on artificial intelligence.

Due to easier understandings and applications, at the initial stage of the load research, statistical techniques are preferable.

Data can be analyzed in the intervals as stored or aggregated upward; i.e., 15-minute demand data can be analyzed as 30-minute or 60-minute demand data.

The sample validation assumes that reasonably complete data is available from sample customers before the beginning of the data analysis. Sample validation can be done by comparison of sample average consumption with the population average consumption for each type of customers and stratum.

Data loss occurs in every study, particularly those involving lots of devices and procedures. Therefore, it is necessary to take missing data into account. In general, if data for a customer is invalid for the time period of interest, the customer is excluded from the analysis. The only exception to this may occur if the analyst determined that excluding the customer will result in a larger error than including an estimate of the specific value. If partial data is available, the analyst must determine guidelines for the amount of missing data that will be tolerated or replaced.

Most load research samples are stratified on the basis of energy consumption. During the course of a survey, some customers commonly migrate from one stratum to another; that is, their consumption increases or decreases significantly such that they are no longer a member of their original stratum. If substantial changes occur in the size, usage, or distribution of customers in the population, redefining stratum boundaries may be an option. However, post-stratification is usually done after a few years of load research or if significant changes are occurring in the total population.

Energy consumption reported from load research surveys is the summation of (15-minute , 30-minute or hourly) demand for the period of interest and is normally reported for calendar month periods. During analysis, the analyst should pay attention of the billing period of energy consumption, and measuring period of interest duo to the load research.

Data analysis is done for defined subclasses/classes of customer and for the system in a different time period (week, month, season) according to type of day (work day, Saturday, Sunday, special day) and different geographic areas specified at the beginning of the load research.

The content of the Load Research Report basically includes:

- Data that could be easily identified and calculated from the collected data in the measuring campaign,
- Estimation of the class demand and system demand at the level of the population,
- Calculations of different factors, and
- Graphics.

On the basis of recorded load data, the following values per customers, subclass and class, and consumption area in the observed time period are identified:

- Maximum load P_{max} (and maximum energy consumption W_{max}) ,
- Minimum load P_{min} (and minimum energy consumption W_{min}),

and determined

- Ratio P_{min}/P_{max},
- Average load P_{sr},
- Energy consumption W.

The typical factors are calculated

- *Load factor,*
- *Coincident factor within the sub-class/class,*
- *Peak occupation factor,*
- *Contribution to the peak factor, and*
- *Coincidence factor.*

The Load Research Report should include following diagrams:

- Load diagram for each customer type at the sample level,
- Normalized load diagram for each customer type,
- Estimated load diagram for each customer type at the population level,
- System (network) load diagram,
- Load duration curves,
- Diagram kWh for each customer type,
- Diagram for each customer type normalized to the global,
- Group kWh diagram.

Using the above-mentioned information from recorded data and statistical techniques, classes and system demand, peak demand and energy at the level of population must be estimated. There is typically no one best estimation procedure.

Mean-per-unit is the simplest and most easily understood estimation method and has a history of acceptance by utility commissions. Therefore, mean-per-unit can be used for the analyses of data collected in the first measurement cycle.

Difference, ratio, and regression methods have a number of advantages and are recognized as appropriate methods in many jurisdictions. All these methods rely on the assumption that demand and energy (or other auxiliary variables) are correlated. It would be good to expand the analysis of data collected in the second measurement cycle by ratio or difference technique.

With adequate sample sizes and no unusual distributions or data anomalies, all the techniques should provide relatively close estimates with acceptable confidence ranges. Estimating

demand using more than one method is an appropriate way to ensure that data problems or poor analysis is not unduly affecting the results.

The analyst must understand applied techniques to avoid errors due to misunderstanding of the technique applied. During the application of the chosen statistical technique, pay attention on correct the use of stratified sampling equations.

Desired precision levels are defined at the time of sample design. It might seem desirable that all demand estimates for all hours of the day and all days of the year should meet the defined criteria. However, the sample precision varies by hour, i.e., the peak hour may have worse precision than during periods of low customer demand. A utility is not required to show that the resulting load curves have this same accuracy but is required to point out any significant deviations from expected accuracy.

In the case of analysis of sampling reliability, energy consumption per customer is an acceptable proxy for demand. If the mean and standard deviation of the sample compare closely with that of the total population, the sample can be considered reliable with respect to those specific parameters. Energy consumption of the sample should correspond closely to the target population consumption (per customer), not only annually but also for each month of the year.

Timing is an important element in reporting load research. Delay can be costly. Both the user and the analyst must understand the time constraints of a load research study and establish realistic deadlines.

Load research reports are prepared for utilities, commissions (regulators), and market agents. Distribution may be limited to the department requesting the information. The reports also contain some information and conclusions that might be interesting for the media and the common citizen.

In future load research, expansion of basic load research reports, including analyses of demand and energy consumption related to temperature, as well as other related variables that mainly depends on utilities, regulators, market agents and customers needs, must be thought about.

Data Transfer

LOAD DATA TRANSFER

Introduction

Transferring load research data between utilities could help the borrowing utility avoid the cost and time to implement a load research study.

Collecting and processing of load research data requires significant dedication of both financial and human resources. Besides, it usually takes a lot of time from the point when the decision is made to the point when the load research data is collected and analysed. Both financial and time requirements make it necessary to have well developed load reasearch patterns . However, in cases when it is not possible to implement load reaserch, the alternative solution is to transfer load research data for a target population that was gatherd by some other company.

Although transferring load research data between utilities can theoretically have many benefits, several obstacles may preclude success. In other words, the factors that make customer loads vary (e.g., appliance saturations or weather) make the transfer of load research data between utilities more difficult.

Particular customer sociodemographic characteristics, weather, type of business activity, and industrial process are some of the factors that influence how specific customer groups use electricity and their resulting load shapes. The importance of these factors in influencing load profiles obviously varies among customer groups (e.g., residential customers are often more weather-sensitive than commercial customers). The following majorfactors to be considered when transferring load data from one utility to another:

- Differences in the saturations of major appliances
- Differences in general economic conditions between the service areas
- Comparisons of demographic data
- Differences in industry types between the service areas
- Differences in the price of electricity and alternate sources of energy
- Differences in climate and the resulting effect on a weather-sensitive load
- Availability of competitive fuels for functions such as space and water heating

These factors pose a problem for load data transfer because the characteristics of the population from which the data were collected (i.e., the donor utility) are often different from the population to which they are being transferred (i.e., the recipient utility). Suppose Utility A wants to borrow some load data on residential customers from Utility B. However, the saturation of electric central air conditioning and electric water heating is much higher in Utility B than in Utility A. Unless some adjustment procedure accounts for those differences, the load data borrowed from Utility B may not provide a useful representation of the load characteristics of the residential customers in Utility A.
To further emphasize this important point, consider the data presented in Figures 1 and Figure 2.
Figure 1 shows the class load shape for the residential class in Territory A of a major utility on the day of its system peak. Figure 2 shows the load curve for the utility's residential class in Territory B on the same day.

Figure 1: Class load shape for the residential class in Territory A

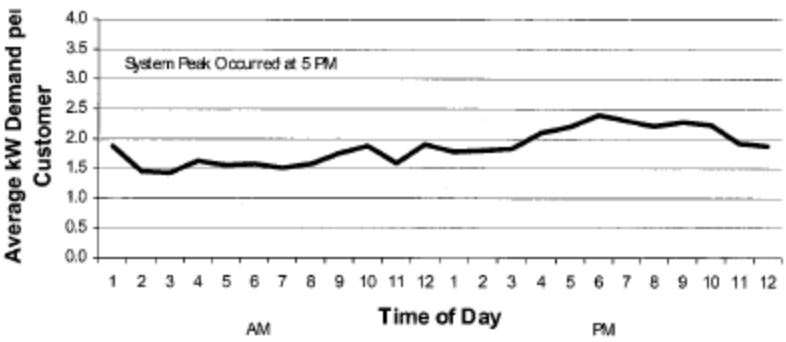
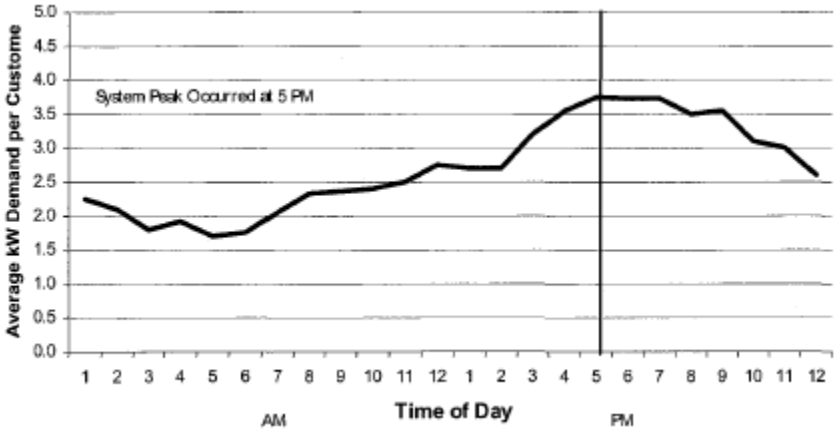


Figure 2: The load curve for the utility's residential class in Territory B



The cooling degree-hours were the same in both jurisdictions. Nevertheless, substantial variation exists between the residential customer load shapes. The appliance saturation survey data presented in Table 1 sheds some light on the reason for the difference. Notice the substantial differences in the saturations of air conditioning between the two groups. Obviously, a transfer of load data between these two groups should adjust for the differences in the appliance stock between the two groups.

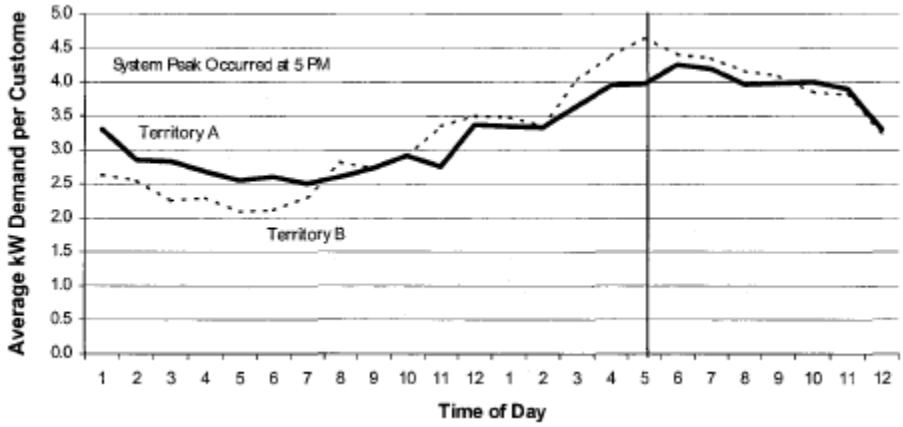
Table 1: Residential Load Survey Selected Appliance Saturations (in percent) by Jurisdictions

	% Customers	
	Territory A	Territory B
Type of Dwelling		
Family house	54,3	85,5
Building	31,9	4,4
Apartment	13,8	9
Mobile home	0	1,1
Type of Air Conditioning		
None	78,8	36,2
Electricity powered	21,2	63,8
Other Appliances		
Electric water heater	8,7	20,6
Dishwasher	34,6	68,6

Electric dryer	33,8	68,1
Freezer	43,4	50,1
Electric range	14,8	53,5

The results od comparison of load shapes for customers with airconditioning on territories A and B are shown on Picture no 3.
These loads shapes, although still different, are considerably more similar to one another than the total residential class load shapes presented in Figures 1 and 2. For this example, customers with central air conditioning possessed similar load shapes on the day of the system peak in two different jurisdictions. This also illustrates that load data transfer is more likely to be successful when both groups have similar weather and sociodemographic characteristics. However, the important factors in determining the differences in load characteristics between groups cannot always be identified.

Figure 3: Residential Customers with Air Conditioning - System Peak Day



Load Curve Data Transfer Method

There are multiple methods that can be used for load data transfer, but for the initial reserach the following ones are suggested.

Mean-Per-Unit/Number of Customers Method

The simplest method of updating load data is to adjust the transferred data to the current number of customers in the population of the other company. If the total estimated demand in the Company B where the load research was carried out is

(1) $T_B = N_B \, kW_B$

N_B - total number of customers in the company B,
kW_B - estimated kilowatt demand per customer for company B

then the updated estimate of total demand for company A would be

(2) $T_A = N_A \, kW_B$

where N_A = total number of customers for company A.

Example 1 illustrates the use of the mean-per-unit method for updating load data.

Example 1 - Mean-Per-Unit Update for Small Commercial Customers

In a load study for the Company B the total number of customers in the class was 25.912 and the average demand at the system peak was 37,6 kW. The estimated total demand was

25.912 customers ·37,6 kW/customers = 974.291 kW.

The total number of customers in the Company A for the same category was 27.273. Updating estimates, the total demand for the category in A, was

27.273 customers · 37,6 kW/customers = 1.025.465 kW

Although this approach is easy to implement, it adjusts the total loads between the two companies only for changes in the number of customers.

The method can be implemented separately for each stratum of a class for which the load study is made. In this case, the effects of a change in the distribution of customers across strata are captured. However, whether the method is applied at the class or stratum level, changes in average customer behavior within a class or stratum are not captured, and no adjustment is made for differing weather patterns or other factors.

Load Factor Method

The load factor method uses the load factor calculated in the company B to transfer the data to the company A. The load factor in the company B is

(3) $$LF_B = kWh_B/(T_B HR_B),$$

where

kWh - total kilowatt-hours for category B,
T_B - total class demand for a specific hour in B
HR_B - number of hours in metering pariod.

Rearranging yields

(4) $$T_B = kWh_B/(LF_B HR_B).$$

To update historic load data, the load factor method substitutes kWh_A with kWh_B and HR_B in equation 4 using load factor LF_B to estimate the total class load in the company A:

(5) $$T_A = kWh_A/(LF_B HR_A),$$

Where

kWh_A = total kilowatt-hours for category A
HR_A = number of hours in the period of consumption metering in A

This method is illustrated in Example 2. With a constant load factor, this method adjusts for changes in weather and customer characteristics to the extent that they are reflected in changes in class sales over time.

Example 2: Load Factor Update for Residential Customers

In a load study for B household sample the duration at time of system peak was 3,2 kW and sample kWh for that month was je 1.284 kWh (the number of hours in the month is 744).

The estimated load factor based on system peak demand was

$$\mathrm{LF} = \frac{1.284\mathrm{kWh}}{3{,}2\mathrm{kW} \cdot 744\mathrm{h}} = 0{,}539$$

The total kWh for category B was 137.862.534 so the total estimated kW for the category was

$$\frac{137.862.534\mathrm{kWh}}{0{,}539 \cdot 744\mathrm{h}} = 343{,}783\mathrm{kW}$$

In the company A, kWh for the household category was 150.378.291. To update the estimated category kW at system peak

$$\frac{150.378.291\mathrm{kWh}}{0{,}539 \cdot 744\mathrm{h}} = 374.993\mathrm{kW}$$

Caution should be used with this method, also, because the load factor often changes over time. For example, suppose in year when metering was done for company B day weather was extreme but the peak month weather was relatively mild. In company A, both the peak-day and the peak-month weather were relatively mild. In this case, applying the load factor method based on data from company B with energy usage from company A will probably overestimate the demand at system peak for company B.

Class-Load Curve Buildup by Stratum

This approach reweights class load data by stratum to reflect the customer distribution in the recipient utility's service territory. This approach possesses the advantage of specifically adjusting the transferred data for differences in the size distribution of customers between the borrower's and the lender's service territories, assuming that the demand per customer within stratum is consistent between the borrower and the lender. The method is similar in concept to the post-stratification method discussed earlier, and the same caveats apply.

If the donor utility's average load per customer for a particular class is calculated by:

(6)
$$\overline{kW} = \sum_{h=1}^{H} W_h \overline{KW_h}$$

where:

W_h = Nh/N weight of stratum h in the donor utility

$\overline{kW_h}$ = average demand per customer in stratum h,

Then the class load curve buildup estimate of average demand per customer in the class in the recipient utility is

(7)
$$\overline{kW_R} = \sum_{h=1}^{H} W_{hR} \overline{KW_h}$$

Where

$W_{hR} = N_{hR}/N_R$ weight of stratum h in the donor utility,
N_{hR} = number of customers in startum h for the observed recipient category,
N_R – total number of customers of the recipent for the observed category,
kWh = average demand per customer in stratum *h* in donor company

A hypothetical example of this technique is shown in Example 3. The class load curve buildup methodology presented here assumes that the average demand or kilowatt per customer within a stratum is constant between the donor and recipient utilities, or is weighted by kWh, not the number of customers. Of course, this assumption may not be valid in all cases.

Example 3: Load Data Transfer Class Buildup by Stratum

Stratum	W_h	$\overline{kW_h}$	$W_h \times \overline{kW_h}$
Donor Utility			
1	0.53	1.2	0.636
2	0.38	2.0	0.760
3	0.09	3.8	0.342
Total			$1.738 = \overline{kW}$
Recipient Utility			
1	0.43	1.2	0.516
2	0.45	2.0	0.900
3	0.12	3.8	0.456
Total			$1.872 = \overline{kW}$

The class load data by stratum from the donor utility are combined with the stratum weights of the recipient utility to produce an estimate of class loads in the recipient utility $\{\overline{kW}_g\}$.

CONCLUSION

It is possible to transfer load research data by using one of the above described methods. When transferring data one should have in mind the adequacy and complexity of the data to be used by the recipient. If it is possible, it is useful for a company to use data from more than one company for development of their own load curve.