TRANSFORMING DUSTY, SELF-SELECTED AUDIT DATA INTO SHINY NEW POPULATION ESTIMATES OF ENERGY USE

Thomas S. Michelman, XENERGY Inc., Burlington, MA
Miriam L.Goldberg, XENERGY Inc., Madison, WI
Andy Loose, Dayton Power & Light, Dayton, OH

Introduction

This paper presents the methodology for leveraging a rich database of self-selected customer audits into estimates of population energy use, EUIs, and saturations. Utilities that have conducted extensive audits with a standardized audit tool should possess data amenable to the described methodology. The methods presented below were utilized for an analysis of Dayton Power and Light Company's (DP&L's) Business and Government (e.g. C&I) customers. The analysis was based primarily on customer data collected through the utility's B&G (Business and Government) Audit program, with supplemental information from other sources.

DP&L initiated this study to develop characteristics of their nonresidential customer segments. The primary objective was to develop better inputs for their forecasting model. In addition, this study provided better information for planning and assessment of customer programs and services. Results of this study have been used by:

- auditors to target customers for special audits for performance contracts;
- the marketing department for segmentation and targeted marketing;
- the evaluation group to confirm basic customer information; and,
- managers to improve the quality of future audits.

This study was designed to take advantage of the extensive data that have been collected at DP&L for the B&G audit program of approximately 5,000 accounts over 5 years. The audit database included detailed information on building structure, equipment, energy usage by end uses, and square footage (so EUIs could be derived at the end-use level). DP&L wanted to use this rich source of data to develop forecast inputs. There were a number of obstacles to overcome in the development process:

- the participants in the audit program were self-selected, not representative;
- DP&L did not trust their SIC on their billing system, therefore no internal building type totals of energy use could be derived; and,
- over the four years of auditing, there had been changes in the data collection and coding system, and therefore cleaning and screening was a major task.

The major steps in the study were the following:

- Data screening and cleaning (sounds obvious and boring, but in actuality it was complex, major, and boring)
- Post-Stratification
- Estimation of Population Estimates

The data sources utilized in this effort included:

- an audit data base of approximately 5000 customers;
- customer billing data; and,
- Dun and Bradstreet data specific to the geographic areas covered by DP&L.

The segments of interest to DP&L were defined by the type of activity taking place in the building. For the audited buildings, the building type was part of the audit record. However, for the general population of business and government customers, building types were not known. Thus, for example, while the energy intensities of large and small office buildings can be determined from the audit data, the total size of DP&L's office sector was not known.

Because of self-selection, it is likely that the proportion of office buildings (used for segmentation) among the audited customers was not equivalent to the ratio of office buildings in the general population. Similarly, at the sub-segment level, it is likely that large and small buildings were disproportionately represented among the audits compared to DP&L's actual population of office buildings.

To reduce the possible bias due to this self-selection, a weighting scheme was used. The weight assigned to each audit was the number of buildings in the population represented by the audited building. These weights were developed from supplemental information on the size of the customer segments in DP&L's service territory. Specifically, the audit data set was post-stratified using Dun and Bradstreet data. Totals and ratios of totals (such as energy and demand per square foot) were calculated using the post-stratification weights. Figure 1 below shows a schematic of the general steps used to derive the population estimates.

The remainder of the this paper will detail the steps which were implemented for DP&L. Also included are insights for improving the process for future studies of this nature.

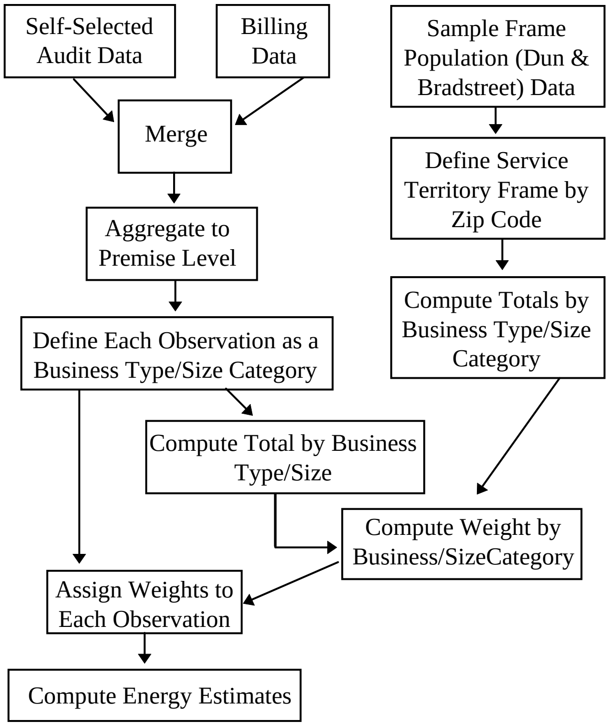
Self-Selected Audit Data
Billing Data
Sample Frame Population (Dun & Bradstreet) Data
Merge
Define Service Territory Frame by Zip Code
Aggregate to Premise Level
Compute Totals by Business Type/Size Category
Define Each Observation as a Business Type/Size Category
Compute Total by Business Type/Size
Compute Weight by Business/SizeCategory
Assign Weights to Each Observation
Compute Energy Estimates

Figure 1

Data Sources

This study is based on analysis of existing data, primarily the DP&L audit data collected on B&G customers

This analysis was based primarily on data from the following sources.

- An audit data base. DP&L has an on-going audit program whose primary benefit to the customers is that they receive energy conservation recommendations. DP&L utilized the Energy Analysis System (EAS) system for their on-site audit program. Similarly, many utilities have access to comparable, on-site data from XENCAP audits, or utility audit programs.
- Customer billing records
- Dun and Bradstreet

These sources are described below.

Audit Data Base

The primary basis of this study is the EAS audit data collected by DP&L. The audit data base included 4,982 audits, conducted over a period from 1991 to 1995. The data resulting from an audit consist of two different types of files of interest.

Firmographic. This file is organized as one observation per audit. The variables of interest include data on:

Address
SIC Code (audit)
Building Type
Floor Space
Employment
Account Number
Operating Schedule Information.

End-Use Equipment Currently Present. There is one file for each of the following categories per audit:

HVAC
Lighting
DHW
Motors
Refrigerators
Miscellaneous.

Each observation is associated with an equipment type (or group of similar equipment). The variables of interest include data on: Equipment description; annual kWh and average annual kW; kWh and kW usage by month by operating status; operating schedule; and, where appropriate, diversity, load factors and usage of other fuels.

Observations

The audit database is very detailed, and clearly such data are expensive to collect. An audit typically includes an initial 2 hour visit, plus a 1 hour return visit to discuss findings with the customer. Between the two visits the auditor checks inputs, analyzes data, and creates a report. On average 4 audits are completed per person week. While expensive to collect, audit information becomes sunk costs, and are free to be used for other purposes (e.g. as input to forecasting models, or for market research). The data that are gathered are raw, but the information that can be derived from these audits is very valuable. No other source of data, with the exception of end-use metering can provide as much detail on end-uses. In fact, the much more expensive end-use metering may not include information on gas or other fuel end-uses. Thus, these audits provide the best and most detailed source of data that an organization is likely ever to have on its customers.

Customer Billing Records

Customer billing records were used primarily to segment industrial customers on the basis of demand level, as described below. In the best scenario, billing data from the most current year(s), would be merged with the audit database to provide the most up-to-date and consistent "snapshot" of population estimates. (The audits for this program spanned the years 1991-1995, and about 1000 B&G audits are completed a year.) For logistical reasons (e.g. not all customers in the audit data base could be matched to a current billing record), annual energy consumption and seasonal peak demand were taken from the audit data base, not from the customer billing records. For those cases where matching was possible, the audit consumption and demand data were found to be reliable. The peak demand billing data were used to confirm the veracity of the audit peak demand data, and were inserted in the analysis database when audit historical peak data were unavailable.

Population Frame Data

To employ the post-stratification methodology (i.e. weighting), as described below, there needs to be a database to post-stratify upon. For this analysis, the database needed to include observations (or aggregation of observations) for every one of the business and government entities in the DP&L service territory. Data collected from Dun and Bradstreet met these prerequisites, and these data served as the basis for post-stratifying the audit data. Dun and Bradstreet provides economic data, including number of businesses, total sales, and total employment by zip code, SIC, and employment size category. DP&L provided XENERGY with a list of zip codes covered by the service territory. XENERGY then obtained the number of businesses, sales, and employment category, totaled across these zip codes, by SIC category.

Observations. While we utilized the D&B derived database called MarketPlace as the basis of the sample frame there were exceptions. D&B has the best information, in cases where organizations need to be credit-checked to perform business. However, the information from the D&B database was weak in some sectors such as Schools, and Colleges (and Universities). The three major problems encountered were:

1. for Schools and Colleges, the audit employment data appeared to be unreliable, often including the number of students rather than the number of employees;
2. a major university was not included (e.g. Wright State); and,
3. For some universities the D&B database had entries only for a handful of departments, not the whole university. (e.g. University of Dayton had five entries, the largest two being the athletic and the economics departments, but in no way comprising the whole campus.)

For these segments we used the number of facilities, rather than total employment, as the basis for weights in these business types. That is, the weight for each school or college cell was simply the ratio of the number of facilities listed in D&B to the number in the audit analysis data set.

Colleges were split into large and small categories based upon D&B employment (not audit employment) for assignment of weights. For this purpose, facility-specific data were purchased from D&B for the colleges. Because of difficulty matching D&B facility listings to those in the audit data base, we still did not apply the employment-based weights to colleges.

In retrospect it would be advisable to substitute another source of sample frame information for the D&B data, when possible for the problem segments This would entail getting another more comprehensive list of sites for a service territory. For colleges and universities, this would be as simple as extracting data from a Barons Guide on colleges and universities for the towns in the service territory.

Another special case was the Wright Patterson Air Force Base (WPAFB) for which DP&L had better population data than D&B. A listing of all buildings at Wright Patterson Air Force Base served as the basis for population when scaling the audited buildings to the entire facility. The list indicated for each building whether an audit had been conducted, and the total floorspace. WPAFB was considered separately from the rest of the DP&L service territory. WPAFB audits were given a weight based on the amount of total square feet/audited square feet. Further, the zip code associated with WPAFB (45433) was not included in the definition of the DP&L service territory used to weight the rest of the business categories.

Audit Data Screening

The primary goal of the audit is to make recommendations of cost-effective energy conservation measures that can be installed by the customer. The auditor collects a vast array of information on energy usage to make such recommendations. While the data are plentiful, they are not structured to address the goal of this analysis. A major effort in this study was to organize the data in a format appropriate for our analysis, and to screen the data for possible anomalies.

In general, a single audit record corresponds to a single building. Each audit record is identified by an audit analysis number. In cases where multiple buildings were audited for a single customer at a single location, XENERGY aggregated the data to the premise level. The premise identification number used as the basis for aggregation is an embedded part of the customer account number.

The aggregation to the premise level was performed for the following reasons.

- Audit data aggregated to the premise more closely matched the Dun & Bradstreet database used for post-stratification.
- The employment variable that was included in the demographic database was more often consistent with a premise level, than an account level, rendering of the database (i.e. premises that had multiple accounts associated with them more often than not had duplicate employment entries across accounts).

Developing the Primary Analysis Data Set

The goal of the initial data processing was to create a database that had one observation per premise, containing demographic information and energy end-use information. The energy end-use information included data on electric (Annual kWh, peak winter and summer period diversified kW) and non-electric energy (annual therms, oil, coal, wood) usage by end use. The final end uses included:

- Cooling (broken out from the HVAC file)
- Heating (broken out from the HVAC file)
- Office Equipment (broken out from the Miscellaneous file)
- Cooking (broken out from the Miscellaneous file)
- Miscellaneous (remainder from the Miscellaneous file)
- Lighting
- DHW
- Motors
- Refrigerators

Office Equipment, Cooking, and Miscellaneous were broken out from the original audit Miscellaneous records. Also included in the end use file was efficiency information for motors and lighting measures.

For each audit, the individual end-use measure data was summed for the building. These data were then merged with the demographic file, which also consists of one observation per building (audit). These two files were then merged. If a premise consisted of more than one building, then the data were further aggregated to the premise level.

Special Data Processing Issues

Needless to say, there were many "opportunities" for creative problem solving during the creation of analysis dataset. These included:

- Finding subset audits containing information on the lighting end use only. These "lighting only" audits were discarded from further analyses.
- Concluding that audit inputs of key variables (building type, employment and square footage) from a handful of auditors were suspect. These audits were discarded from further analyses.
- Trying to discern which version of the audit software had been used to input audit results. This was a major issue as the same building type value would resolve to a different building type description depending upon the version of the audit software.
- Merging data from the audit and billing databases, when the billing information system database had been completely changed (including the linking ID) during the course of the data collection period.
- Imputation of square footage or employment when either were missing.

Review of Energy and Demand Intensities

The above "opportunities" made it clear that energy and demand intensities needed to be reviewed closely. Energy and demand intensities were computed for every end use for every building. The energy intensity is the annual energy consumption for the end use divided by the building's floorspace (e.g. kWh per square foot). The demand intensity is the peak electricity demand divided by the building's floorspace, in kW per square foot. For natural gas, whole-building energy intensity was computed, in therms per square foot, but no end-use intensities were examined. Determining end-use intensities for gas were not a goal of this study.

For each building type, we summarized the distribution of each energy and demand intensity across buildings of that type. The distribution statistics reviewed were minimum, first quartile, median, third quartile, and maximum. Any building that had an end-use intensity greater than 10 times the third quartile was flagged as an anomaly, and referred for closer examination. Examination of such cases identified the following common situations.

- Recorded floorspace apparently in error, too low, based on building name, or employment level. For example, we found a school with 1,000 square feet.
- High intensity apparently correct, based on type of activity and equipment. For example, we found several well houses or pumping stations with small floorspace and large amounts of pumping equipment, accounting for extreme motors EUI's.
- Building type apparently in error, as a result of incorrect coding scheme being applied.
- Uncertain. Initial review of the information did not indicate whether or not floorspace or equipment were likely to be in error.

Anomalous EUI's in these four categories were handled as follows:

- If the floorspace appears to be in error, but employment data are available, impute floorspace based on the employment, and the average square feet per employee among other buildings of that type. Keep the building with the imputed floorspace, and set the imputation flag. If employment data are not available, drop the building.
- If the high intensity appears to be correct, keep the building without change.
- If the building type appears to be in error, correct it if possible. If not, drop the building.
- Small uncertain buildings, which would have little effect on the final estimates, were left as is, because final estimates are implicitly weighted by floorspace. Large uncertain buildings were reviewed by DP&L.

As an additional check, the largest 10% of buildings by square footage, accounting for over 40% of the total square footage were reviewed by DP&L staff familiar with the audit procedures and specific sites, whether or not there were concerns.

Post-Stratification

Because the audited customers are not a random sample of customers, a post-stratification scheme (weighting) needs to be applied to project estimates of the population's energy use and saturation. The idea of the post-stratification scheme is to classify each audit according to size and building type, then assign a weight to each size-type cell based on the number of such buildings in the population. For example, if the population has 300 medium schools, and the available audit sample has 15, we would assign a weight of 300/15 = 20 to each of the medium-sized schools in the audit sample.

The difficulty with applying this method is that DP&L's population of B&G customers are not coded by building type. SIC code information is attached to the customer billing records. However, DP&L staff indicated that this information is not reliable, and recommended that it not be used for this analysis.

XENERGY considered two approaches to developing post-stratification weights. One was based on fielding a survey to a large sample from the general B&G population. The other made use of existing data available from Dun and Bradstreet. The potential improvement in accuracy from fielding a large survey had the following drawbacks:

- it would be expensive and time consuming;
- it would put a burden on DP&L's customers;
- it would only be a survey of a sample of customers; and,
- we would still be left with questions of possible non-response bias.

As noted above, the post-stratification scheme adopted relied on the D&B data.

Basic Post-Stratification Approach

Dun and Bradstreet (D&B) data are publicly available at the aggregate level at no marginal cost. That is, the only cost associated with using these data is the cost of the analytic effort to query the D&B system to obtain the desired summaries. D&B data available for each business include employment, 4-digit SIC code, and zip code. They do not include floorspace.

We obtained employment totals by SIC group and employment size categories, summed over all zip codes in DP&L's service territory. The weighting procedure then proceeded as follows.

1. Define business types as groups of SIC's, roughly corresponding to the building type categories defined for this analysis.
2. Define employment size categories, separately for each business type. The size categories were set to divide each business type into three groups of roughly equal total employment, based on the D&B data. Exceptions were made to ensure that the top size category included at least four or five audits.
3. For each business-type/size category *c*, sum up the total number of employees E_{AUDc} in audited sites, from the audit data.
4. For each business-type/size category *c*, obtain the total number of employees E_{DBc} in DP&L zip codes, from D&B.
5. For each business/size category *c*, calculate the weight as $R_c = E_{DBc}/E_{AUDc}$.
6. Assign the weight $W_j = R_c$ for each audit *j* in business/size category *c*.

An alternative at Step 5 would be to base the weights for each business/size category not on total employment but simply on number of cases. The difficulty with this approach is that the units being counted are not the same in D&B as in the audit system. D&B has a record for each business at each location. The audit database has a record for each building, which we have aggregated to a record for each site. What was ultimately of interest to DP&L is a count based on the number of accounts, which is the basis of records in the customer information system.

This difference in what is the unit of observation--premises or businesses--does cause some difficulty in assigning audits or premises to post-stratification cells. For example, a large office building with many small tenants would be assigned to a "large" size cell based on the audit data. However, in the D&B data, this building would show up as several small businesses; the total employment would appear in the "small" category.

A general difficulty with any post-stratification scheme is that it may lead to extremely high weights for certain cells if the audit data happen to have very few cases in those cells, but the population is large. This possibility can lead to high-variance estimates. That is, the estimates could be very sensitive to a few cases with extreme weights. To limit the potential for extreme weights resulting in unstable estimates, we specified a minimum sample size of four for each cell. (An exception was made for colleges, where the three largest in the audit data base were substantially larger than the others.) We also reviewed the weights produced by the procedure outlined above, and determined that there were no excessively large weights associated with large customers, which could cause unstable estimates.

The post stratification weights were based upon the following business categories:

- Agricultural
- Amusement & Recreation
- Apartment
- Auto Related
- Business Service
- College-Small
- College-Large
- Electric, Gas, Sanitary
- Food Store
- Furniture Stores
- Government
- Health Services
- Hotel/Motel
- Industrial, Machinery & Equipment
- Industrial
- Membership Organizations
- Miscellaneous
- Office
- Personal Service
- Restaurant
- Retail
- Schools
- Transportation
- Weight Patterson Air Force Base
- Wholesale

Estimation

The weights are computed separately for each stratification cell, and assigned to each audit in that cell. A cell is defined by the post-stratification business type and size category.

Estimates for this study are computed not by stratification cell, but by segment. Although some of the segment names are the same as some of the business type names, an audit in a particular business type post-stratification category would not necessarily be in the building type of the same name, and vice versa. For example, an audited building with an SIC designation of Industrial might have a building type Office or Warehouse.

Once the weights are assigned to each audit on the basis of the business type/size post-stratification, the post-stratification cells are not used in the remainder of the analysis. Instead, estimates are computed by reporting segments, which are building types.

Computation of Weighted Totals

The weights W_j assigned to each audit *j* are used to compute each aggregate estimate for each segment of inter est. Specifically, for any variable *x*, the total of *x* over the population in a given segment *S* is estimated as:

$$X_S = \sum_{jeS} x_j W_j \cdot$$

For example, the total cooling electricity consumption in office buildings was computed as:

$$ACkWh_{OFFICE} = [\Sigma_{j\varepsilon OFFICE} (ACkWh_j)(W_j)].$$

where $ACkWh_j$ is the annual electricity for cooling for premise *j*, from the EAS data.

Likewise, the total floorspace of office buildings is given by

$$SQFT_{OFFICE} = [\Sigma_{j\varepsilon OFFICE} (SQFT_j)(W_j)].$$

where $SQFT_j$ is the floorspace of premise *j*, from the EAS data.

End-Use Intensities

End-Use intensities (EUI) are the ratios of total energy or demand in a segment to the total floorspace in the segment. For example, the cooling EUI for office buildings $ACEUI_{OFFICE}$ is

$$\begin{aligned} ACEUI_{OFFICE} &= ACkWh_{OFFICE} \,/\, SQFT_{OFFICE} \\ &= \left[\sum_{j\varepsilon OFFICE} (ACkWh_j)(W_j)\right] / \\ &\quad \left[\sum_{j\varepsilon OFFICE} (SQFT_j)(W_j)\right]. \end{aligned}$$

(Note that the weights W_j, which appear in both the numerator and denominator of the EUI equation do not cancel out, because they are not constant across different audits *j* in the summations.)

Fuel Shares

The fuel share for a fuel *f* and end use *u* is the proportion of the floorspace in a segment *b* contained in buildings that use that fuel for that end use.

$$FS_{buf} = SQFT_{buf}/SQFT_{b.}$$

Comparison of Sector Weighted Total with DP&L Total

The post-stratification method based on D&B data was used because we did not know the number of DP&L customers in each segment/size category. However, we did know the total consumption for DP&L's B&G sector. As a check on the overall weighting procedure, we compared this known total with the weighted total computed from the EAS data as described above. The weighting expands the audit data according to the 1996 customer profile, as described above. We estimated that DP&L B&G total for 1996 by increasing the 1995 total by 3.4 percent, the average annual rate of increase from 1993 to 1995. The audit weighted total was 10.7% higher than the adjusted DP&L total. The result indicates moderately good agreement of the weighted total with the known DP&L B&G total.

Conclusions

This study demonstrated that substantial information about customers can be developed by mining existing data sets collected for other purposes. A key to the development of the estimates presented here was the post-stratification using Dun and Bradstreet data. This step was necessary because of the low confidence DP&L staff had in the SIC classifications of their customers. This is a problem many utilities share.

The post-stratification scheme worked well. The results for the various segments were reasonable. This judgment is based on our experience with similar studies, as well as comparisons with national data. We compared the intensities and end-use fuel shares for the different segments with data from the Energy Information Administration's Commercial Buildings Energy Consumption Survey for the East North Central Census Division, and found no major anomalies.

While the methods developed here were effective, and provided good quality results with valuable detail, the process is not without costs. Considerable time was spent resolving inconsistencies and developing a clear understanding of the data's interpretation and limitations. In this process, the involvement of DP&L staff who were intimately familiar with the data collection process and system was critical.

As some final "sound bite" recommendations:

- Don't survey your customers a second time when valuable information is already available.
- Leverage what you have. While data from these audits was not perfect, it was much better than the alternative, nothing.
- Consider other ways of integrating data available to you within the utility. Data gathered from monthly billing databases or automatic meter readers can be incredibly valuable in triangulating the usage estimated from audits.
- Understand the benefits and limitations of using audited data. Gauge what percentage of the load you have covered in your audits. If you have only a few customers (or a small fraction of the square footage) for a certain building type keep in mind that you should be less confident in your estimates.
- If you are going to try to transform dusty, self-selected audits into shiny new population estimates, someone who was intimately involved in the audits must continue to be involved in the analysis to resolve the many issues that are bound to arise.