WHEN TRUST MATTERS

Are More Data Better?

Sampling Versus Census Analysis for Class Load Estimation

Curt D Puckett, Vice President, DNV Energy Insights, Key Initiatives
Kevin Cracknell, Principal Consultant

24 March 2021

First a Commercial Break…

AEIC Advanced Applications in Load Research

Scheduled for early November – stay tuned for specific dates/times

Anticipated topics include:

- EV Identification
- PV Sampling & VEE
- Storage
- Non-Wire Alternatives
- Segmentation & Clustering

Today's Key Takeaways

- A well-structured and properly allocated sample provides accurate result
- A sample allows for a higher degree of "tender loving care", i.e., validation, editing and estimation
- A sample reduces analytical costs and processing times
- A "super" sample provides a suitable analytical framework for additional analyses

Commission Staff Position

- *"Sampling is unnecessary when the Company's AMI system can provide nearly all the actual load data, a benefit for which the Company's ratepayers are paying. With the Company having the ability to utilize accurate load data through its fully deployed AMI infrastructure, it should be required to use accurate load data to allocate costs and develop rates in its next rate case."*

Motivation

- ***The last sentence implies that the Company's historical and current analysis approach that uses sampling is not "accurate" for the purpose of developing class load profiles for use in the cost-of-service analysis…***

DNV

Commission Directive

- *Staff disagreed with the use of sampling to build load profiles due and suggested that the Company use load data from all of its AMI meters in the development of its load profiles in its next rate case.*

Company's Position

- Sure, we will instruct our consultant to process the full complement of AMI data to assess the benefits derived from simply "adding up" the near census sample,
- Oh, and while they are at it, we will have them assess the increase in accuracy and identify any associated challenges encountered
- I mean, is it really just a matter of adding up the intervals – some people think so…

DNV

Going Back to our Roots

- PURPA (1978)
  - Effectively required the utility industry to develop load research programs as a basis for cost-of-service filings with FERC and state public utility commissions.
  - Required utilities collect load research data, such as daily kilowatt (kW) demand load curves, for large utility rate classes, specifically those classes that were at least 10% of the overall sales of the utility.
  - Helped establish confidence (90%) and relative precision (±10%) requirements on estimates of selected cost allocation factors, like coincident (system) peak demand.

Other Motivations

- Before AMI
  - Cost per sample point was high – could run in thousands ($5,000+) per sample point
  - Cost of load research was effectively proportional to the required sample size
  - Data processing was more limited and restricted – governed by IT using main frame computers
  - Efficient sampling and estimation methods allowed for better precision with smaller sample sizes
  - Even with AMI there is a cost per sample point for data management and data validation

DNV

Sample Size Requirements

- Sample Size for Stratified Ratio Estimation

$$n \approx \left(\frac{z \ er}{D} \right)^2$$

- Where,

$$er = \frac{\sqrt{\sum_{i=1}^{n} w_i \left(e_i^2 \big/ x_i^{\gamma} \right) \sum_{i=1}^{n} w_i x_i^{\gamma}}}{\sum_{i=1}^{n} w_i y_i}$$

Classic Load Research Residential @ 90%/±10%

- D = Desired Relative Precision, i.e., ±10%
- er = error ratio for key attribute of interest, e.g., system peak demand (er=1.0)
- z = standard normal deviate @90% confidence z=1.645
- n = Sample Size = $(1.645*1.0/0.1)^2$ = 271
- Adjusting for potential data loss of 20% yields a final sample of 271/0.8 ≈ 340
- So, a “good” stratified sample of n=340 should be sufficient but they want us to analyze ≈ 1.2 million customers

Data Available from the Original Sample Design

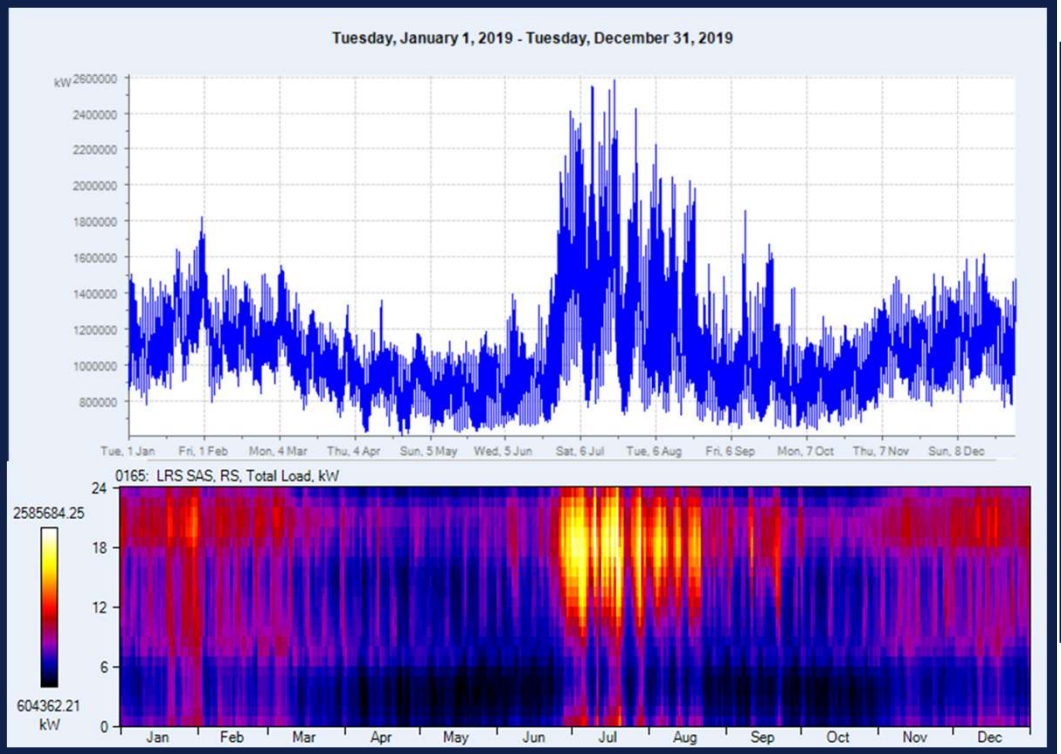
- The table presents the original sample design used in the 2019 analysis
- A five strata design was used
- A large robust sample of 3,073 sample points were included in the final analysis
- The “design” precision at 90% confidence was estimated to be better than ±3%

Residential Class

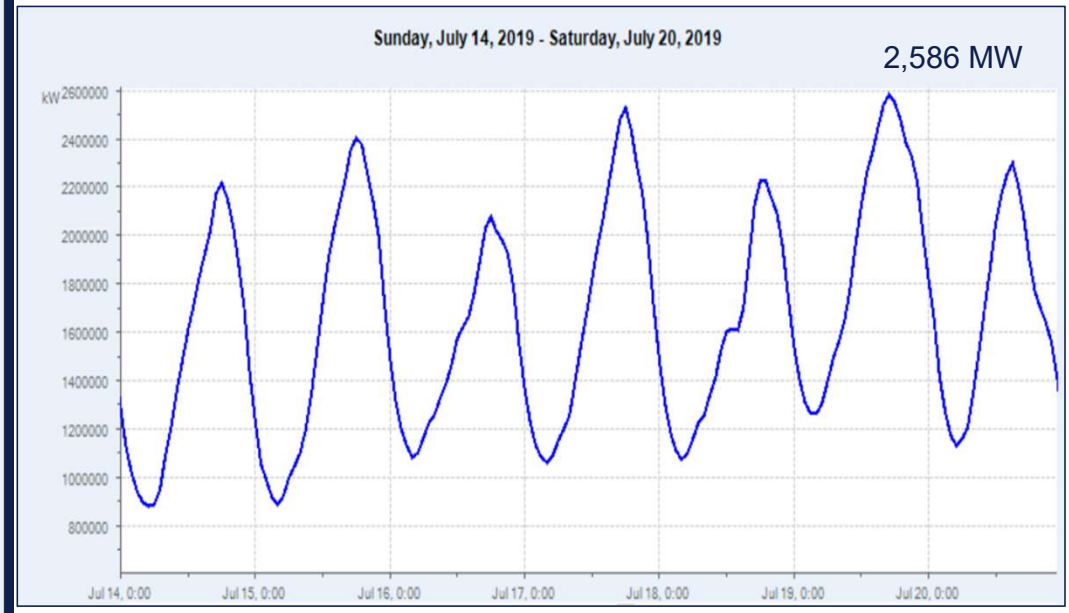
Stratum	Maximum Annual Use (kWh)	**Population Count (N)**	**Population Total Use (kWh)**	Sample Size (n)	**Case Weight (N/n)**
1	0-5,448	463,496	1,498,654,092	771	601.16
2	5,449-7,770	257,202	1,690,118,630	639	402.51
3	7,771-10,326	200,878	1,797,192,264	616	326.10
4	10,327-14,496	158,204	1,908,258,106	552	286.60
5	>14,496	101,231	2,177,083,482	495	204.51
Totals		1,181,012	9,071,306,574	3,073	

DNV

Class Profile

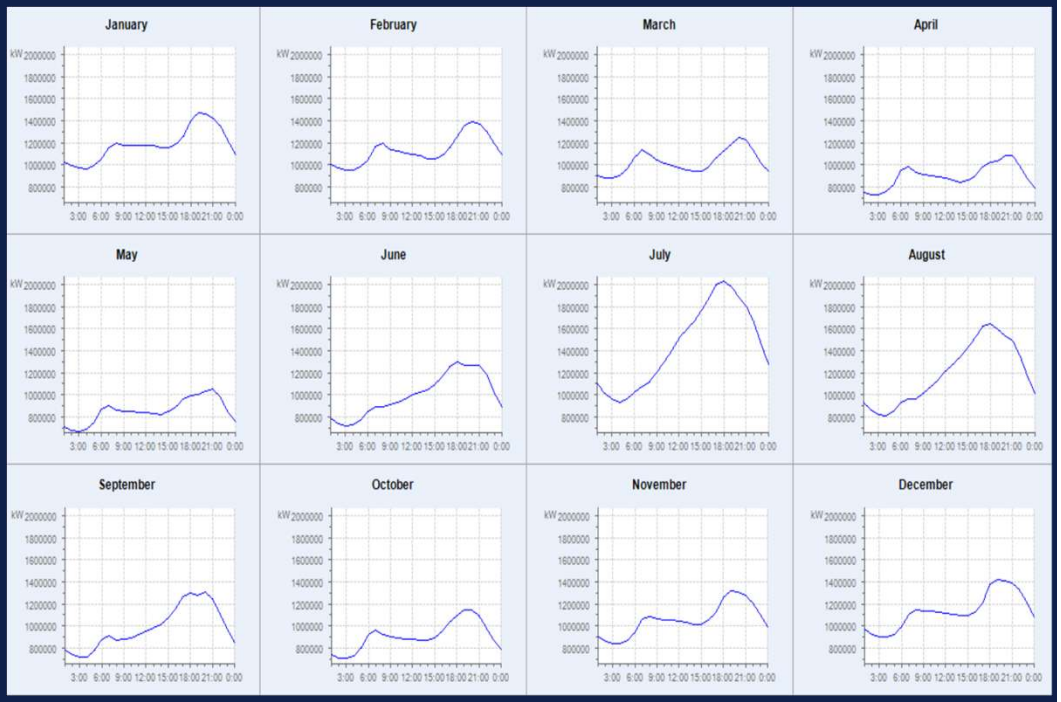


Peak Week



The achieved precision at the class peak was ±2.2%

Average Weekday By Month



Accuracy of Current Sample

Achieved Precision

Updated Availability Using Full Complement

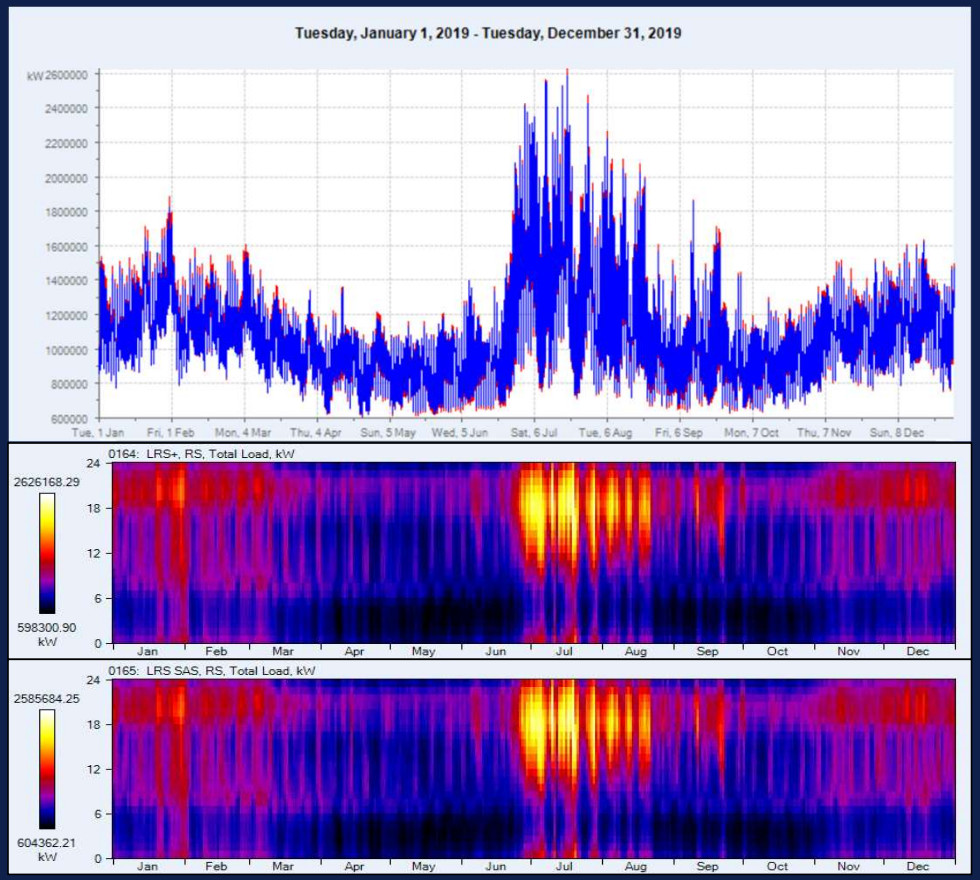
- Not all customers from the population are available to include in the analysis which requires some degree of post-stratification and extrapolation
- Case weights vary from 1.02 to 1.04
- Sample size is now 1.146 Million customers
- 97% of the customers are included in the analysis
- How different to we expect the load profiles to be?

Residential Class

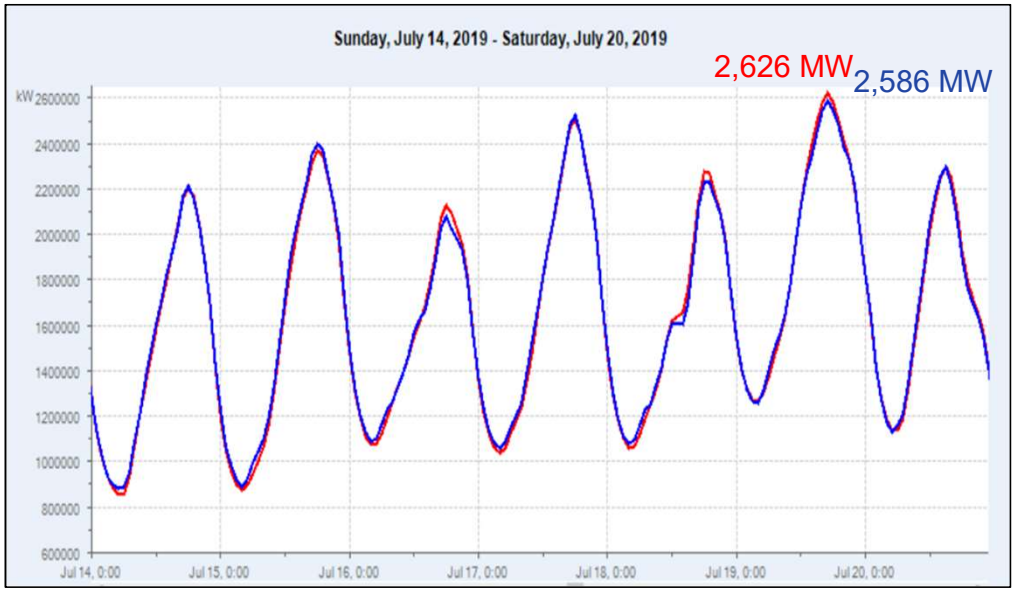
Stratum	Population Count (N)	Population Total Use (kWh)	Stratum Cut Point Max Use (kWh)	Sample Size (n)	Case Weight (N/n)
1	463,496	1,498,654,092	5,448	447,787	1.035
2	257,202	1,690,118,630	7,769	250,991	1.025
3	200,878	1,797,192,264	10,325	196,008	1.025
4	158,204	1,908,258,106	14,495	154,072	1.027
5	101,231	2,177,083,482	711,630	97,751	1.036
Totals	1,181,012	9,071,306,574		1,146,610	

DNV

Class Profile Compare

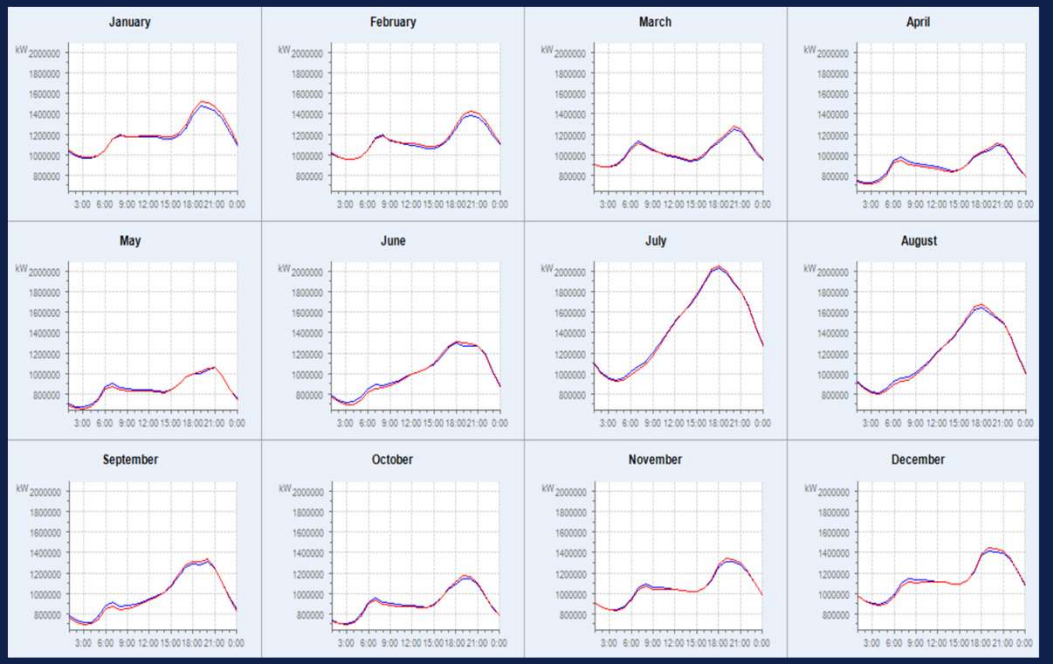


Peak Week Compare



The “actual” peak was calculated to be 2,626 MW versus 2,586 MW or 1.5% higher – this is well within the confidence interval of the original estimate

Average Weekday by Month



Graphical Comparison

- There is no discernable difference in the average weekday load profiles by month.
- Each month shows an identical pattern

Challenges & Lessons

- Data Extraction
  - Querying and extracting 1.2 million meters is computing resource intensive (processor & memory)
  - Running the extraction slows down server, had to be run in off hours
  - Extracted files require a storage drive with substantial space, over 700 GBs of data
- File Transfer
  - Time intensive – uploading and downloading nearly a terabyte of data takes significant time
- Pre-Processing (Data Transformation)
  - System configuration – handling data of this size required setting up optimized spark cluster
  - Time intensive – additional time needed for formatting data into the fit and form required for analysis
- Data Loading
  - Expensive – cloud services charge based on system size. A system with adequate resources to handle 750 GB of data and run analytics increases costs dramatically
  - System configuration & optimization – Required a large scale optimized relational database to handle IOps (input/output per second) requirements

DNV

Challenges & Lessons (cont.)

- VEE
  - Even AMI data needs some level of tender loving care to ensure that it is analytically ready for “prime time”
  - QC costs are proportional to count of data streams used & additional time/expense to quality check 1.2 million customers is not cost-justified vs. “large” sample of (e.g.) 3,000
  - To do it right requires substantial resources or do you make the decision to simply cut and run
  - Billing anomalies will still occur with AMI

- Analytics
  - System architecture – Develop tool with large scale data in mind
  - Many off the shelf analytical software programs are not built to handle big data (terabytes), at least not without substantial hardware/supplemental software investment
  - The “old” cost of metering may simply be offset by the IT infrastructure required to handle this new data asset
- Cloud Resourcing
  - Given the latest technology you can spin up and spin down resources as needed
  - Cloud flexibility is much easier to do, and some internal IT departments may not be used to this more flexible system dynamic

DNV

Conclusions

- A well-stratified and properly allocated sample develops **accurate estimates** of class loads
- Sufficiently accurate for use in the cost-of-service and other utility planning processes
- Sample allows for reduced data management time, increased VEE, and low analysis costs
- Data capture needs to meet the analysis objectives, simply processing more data because it is there may be suboptimal

- Education is required so that Commission staff understand the fundamentals of sampling
- AMI data is empowering, and we encourage utilities with access to unlimited AMI data to use a “super sample” strategy
- A well-designed, large sample encompassing thousands of sample points ensures a robust data stream for the data scientists

DNV

Moving Towards a Super Sample Strategy

- Updated the new sample design to include a significant larger sample to help alleviate Commission staff concerns
- Increases accuracy – actually achieved a precision for 90% of all hours at or below ±0.5%
- Reduces the processing time and data extraction burden of a census analysis
- Provides data scientists more “clean” data for other analytics allowing drill down analysis with meaningful sample sizes

Stratum	Population Count (N)	Population Total Use (kWh)	Stratum Cut Point Max Use (kWh)	Sample Size (n)	Case Weight (N/n)
1	463,496	1,498,654,092	5,448	10,000	46.350
2	257,202	1,690,118,630	7,769	10,000	25.720
3	200,878	1,797,192,264	10,325	10,000	20.088
4	158,204	1,908,258,106	14,495	10,000	15.820
5	101,231	2,177,083,482	711,630	10,000	10.123
Totals	1,181,012	9,071,306,574		50,000	

DNV

WHEN TRUST MATTERS

Thank You

curt.puckett@dnv.com
+1-517-898-7078

www.dnv.com

DNV