http : //www.cigre.org

CIGRE US National Committee
2018 Grid of the Future Symposium

Residential Customer Clustering at Dominion Energy: Towards Big Data Analytics

M. S. MODARRESI [1, 2], E. HALL [2], P. MARKHAM [2], P. IRELAND [2], K. THOMAS [2]
[1]Texas A&M University
[2]Dominion Energy
USA

SUMMARY

Severe cold weather events in recent years have led to the sudden failure of multiple step-down transformers in Dominion Energy's service territory. Investigations have revealed that the transformers were sized based upon assumptions about residential dwelling heating type that was incorrect. The addition of roughly 400,000 smart meters to the Dominion Energy system has created new opportunities for obtaining deeper insight into how energy is used by our customers. Data analytics on the customer's response to daily changes such as the time of day and temperature can reveal valuable information through supervised and unsupervised learning algorithms.

This paper is one of our first attempts to use unsupervised machine learning algorithms to cluster customers' behavior into different groups. We will show how unsupervised learning algorithms can be used in practice and how the required data needed can be reduced to smaller datasets.

KEYWORDS

AMI, clustering, load forecasting, k-means, data analytics, machine learning

sadegh.modarresi@tamu.edu

INTRODUCTION

In late December of 2017, a severe cold wave affected the Central and Eastern United States and Canada, leading to record low temperatures and snowfall that lasted through mid-January of 2018 [1]. During some of the most extreme periods of low temperature, an unusually large number of distribution step-down transformers failed within Dominion Energy's service territory, leading to power outages for hundreds of customers. Further examination has revealed that the transformers were sized based upon assumptions about residential dwelling heating type that did not take into account conversions from gas to electric heat. In addition, energy usage patterns were assumed to be the same for all customers, which has important implications for estimating the peak load experienced by the transformers. These events enhanced the Dominion Energy motivation to have a higher resolution picture of residential customer's consumption.

Beginning in 2012, Dominion began deploying interval billing meters throughout its service territory as part of its Advanced Metering Infrastructure (AMI) implementation. These meters provide 30-minute readings of both real (kWh) and reactive (kVARh) energy, which are imported into our meter data management (MDM) system. Sixty-six thousands of the nearly 400,000 interval meters' usage records were retained for billing or various survey and engineering studies. The remainder of the readings were discarded.

As of December 1, 2017, Dominion began storing all of the 400K interval meter records in our new Big Data platform. Coincidentally, under the recently passed Grid Transformation and Security Act, Dominion plans to replace all remaining meters over the next six years. This means customers' 30-minute reads will be available on a daily basis giving Dominion a huge opportunity to derive benefits from this data. The granularity of interval meter data provides numerous opportunities for understanding each customer's energy consumption. One example of this would be the ability to infer a customer's heating type based upon their energy usage. Another is the load factor (i.e., the ratio of the peak load to the average load), which is of particular interest to distribution system planners.

Beyond heating types, AMI data clustering enables the utility company to bypass estimations about each individual household while clustering customers into groups with the same behavior. In other words, instead of "estimating" if a customer uses an efficient electric heater or not, which is subject to change from time to time, we cluster a set of customers with similar behavior into groups. The clustering algorithm is unsupervised since we do not define the borders of these clusters. In this paper, we used one of the unsupervised learning algorithms for this purpose.

The rest of the paper is organized as follows: In section II, we describe a supervised customer classification method mapping AMI measurements to the heating type used by residential dwellings. In section III we introduce the unsupervised learning algorithm to cluster customers into separate groups based upon their consumption response to changes in the temperature during a chunk of hours of the day. Conclusion and future work will be presented in section IV.

Section II-SUPERVISED LEARNING: HEATING TYPE CLASSIFICATION

It should come as no surprise that different heating technologies exhibit unique electrical usage patterns in response to changes in temperature. For example, a gas furnace would not show much variation as temperatures fall since the blower motor does not operate much more often than when temperatures are more moderate. An air-source heat pump would show gradually increasing energy usage as the outdoor temperature drops, with a sharp increase observed once the temperature falls below 35 °F and its coefficient of performance (COP) decreases. Resistive heating behaves in a somewhat similar fashion, though its usage will be higher at moderately cool temperatures. Given these differences, it is important to know the heating type in use at each residence so that the peak load can be more accurately predicted.

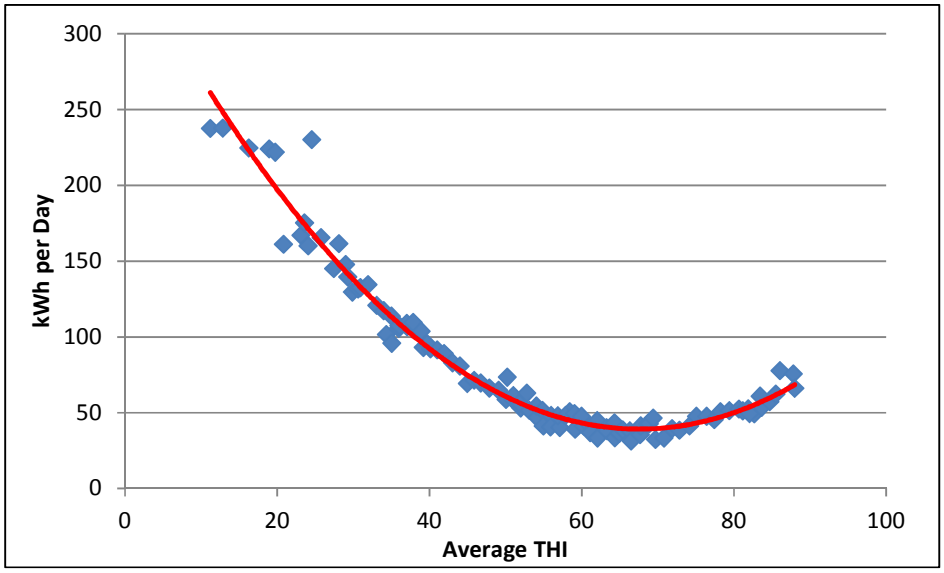

Figure 1 - Benchmark for slope circulation.

The process used to determine the heating type for interval meters is as follows:

1. Determine the customer's baseline energy usage, i.e., the energy consumed during times of the year when the heating system is not used. In this case, April and October are the "shoulder" months, and the bottom quartile is used to determine the average kWh usage per day. It is also important to eliminate days in these months where the outdoor temperatures are above 70 degrees or lower than 50 degrees as some heating and cooling could be taking place.
2. Calculate the winter energy usage per day using the 75th percentile of their daily usage in the months of December, January, and February. Here again, we are only interested in those days where the outside temperatures are below 40° F This ensures that if space heating is taking place that we were able to isolate the usage increase over the baseline period.
3. Calculate the ratio of the winter daily usage to the baseline usage.
4. Fit a regression line to the dependence of energy usage on the squared average temperature-humidity index (THI) for the heating and shoulder seasons. This results in the slope, intercept, and correlation coefficient (r) for the model.
5. Apply the following logic to determine the heating type:
    a. If HeatRatio < 1.0 ➔ No heat
    b. If 1.0 ≤ HeatRatio < 1.5 ➔ Gas heat, type 1
    c. If Slope > 0 ➔ Gas heat, type 2
    d. If (HeatRatio ≥4.0 and Slope < 0 and r^2 ≥ 0.80) OR (➔ Electric heat, type 1
    e. If 3.0 ≤ HeatRatio < 4.0 and Slope < 0 ➔ Electric heat, type 2
    f. If 1.5 ≤ HeatRatio < 3.0 and Slope < 0 ➔ Electric heat, type 3
    g. Else, Unknown

Heating or cooling types can be used as an indication of the consumption sensitivity to temperature changes in the system. There is a significant complication to this approach, however, which is the sheer volume of data that must be processed. For our company, 400,000 interval meters returning 48 measurements per day results in 576 million records per month and almost 7 billion records per year. When AMI metering is fully deployed system-wide, this figure would blossom to 50 billion records per year. Performing this type of analysis using a traditional relational database would take days, if not weeks. Given that the number of interval meters is only going to increase in the future, a better solution was needed. Luckily, Dominion Energy's Information Technology group has recently implemented a Big Data environment intended for this type of problem. The system, which uses Hortonworks Db2 Big

SQL platform, consists of 14 nodes (4 masters, 8 data, and 2 edges) with a total of 172 TB of storage and 4.2 TB of RAM. Access is provided through a web-based RStudio environment using the Apache Spark open-source cluster computing framework.

The algorithm was implemented in the R programming language and executed using the Big Data environment. Total execution time for all meters was less than an hour, a significant improvement over using a relational database. In the next section, we discuss how this index helped in testing the accuracy of k-means clustering as an unsupervised learning algorithm.

Section III- UNSUPERVISED LEARNING: CUSTOMER CLUSTERING AND VALIDATION

Clustering is an optimization problem. It minimizes the dissimilarity between customers by minimizing the Euclidean distance between data points. If no constraints are defined, the algorithm will put each data point into one cluster and then the distance between clusters would be zero. However, no information can be gained from such an optimization. Therefore, there is a need to define constraints for the clustering algorithm.

In the k-means clustering algorithm [2,3], k is used as an upper bound on the number of clusters. Due to the convexity of the objective function, the algorithm always returns exactly k clusters. To cluster customers into different sets with similar behavior, first, we need to know the number of these clusters. However, in most cases, this number is not pre-defined.

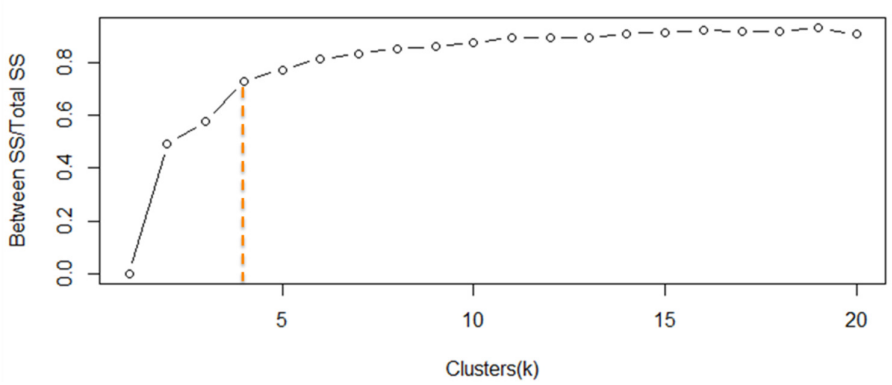

Figure 2 - State of the art approach to find the number of clusters in the k-means algorithm.

In our simulation, we considered days with temperature below 45 degrees for 45 days of Winter 2017-18. To eliminate the different consumption difference between customers, we grouped morning peak hours together using average consumption and temperature-humidity index for these hours as input. The state of the art approach is illustrated in Fig. 2. One starts with a minimum number of clusters, k=1 and progressively increases k while observing the clustering density index (i.e., the ratio of the in-cluster sum of squares to the total sum of squares). At some value of k, the clustering density index stops improving, thus it is chosen as k for future cluster analysis. Figure 1 shows the clustering of customers in a particular city. The suggested number of clusters is four in this case.

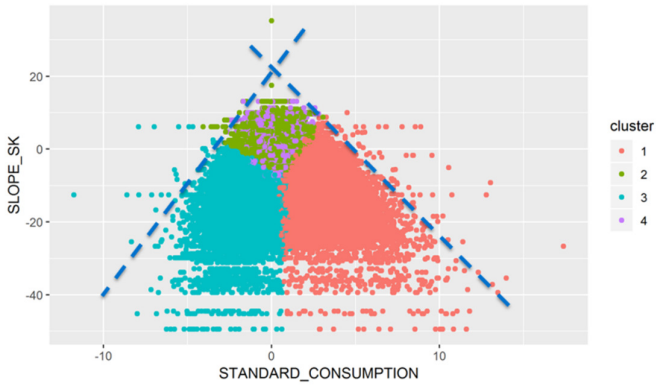

Figure 3 - Customer clustering results considering slope as one of the input parameters.

Figure 3 shows the customer clustering results considering AMI consumption data, the temperature for the ZIP code and regression curve slope. As expected, consumption of customers with positive Slope or slightly negative slope, have a distinct behavior compared to customers with a negative slope in the cold days we considered in this presentation. Also, we might like to see some point with outlier behavior. For instance, the green points on the top or the points outside the dashed line in Figure 3 are possible candidates to be looked into. Figure 4 shows the clustering results from different attributes. On the right, the red cluster shows data points with high sensitivity to cold temperatures, while purple data-points show virtually no sensitivity to changes in temperature which confirms the results in Figure 3 as well.

To show how the k-means algorithm can cluster customers in a similar fashion with and without household details, we first use the slope obtained from the previous section and cluster customers with three attributes, then we drop the slope attribute and compare the results again. If the results are visually comparable, we can claim that there might be no need for a supervised input to the algorithm and customers can be clustered by only using daily AMI and temperature data.

Figure 4 - Customer clustering results using slope.

Figures 5 and 6 show compare the clustering both with and without considering the slop attribute. In Figure 5 on the left, we performed the clustering without using slope, then we added existing slope to the clustered data to visually show the difference between the two, while the right figure incorporates slope into the clustering. Figure 6 tries to deliver the same message using a different attribute of data. As can be seen, the breaking point between the clusters is the same for these approaches. Clustering results are also basically the same aside from some minor differences.

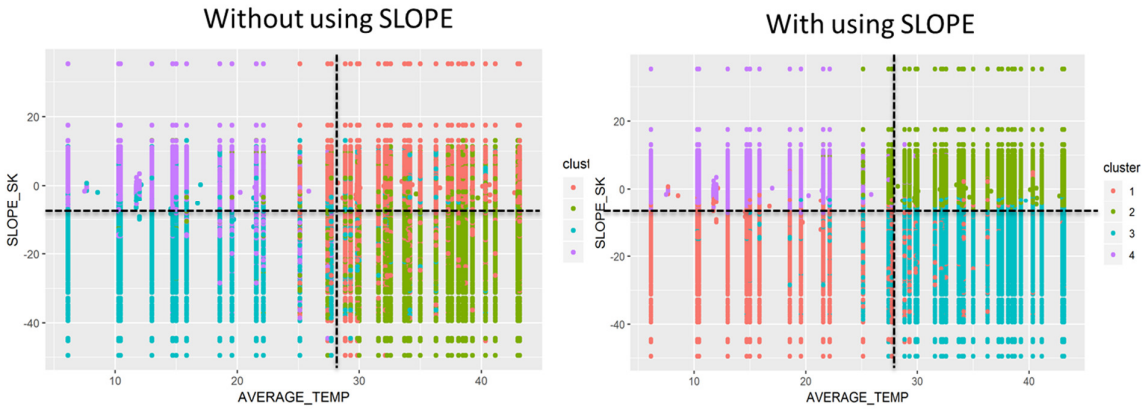

Figure 5 - Results comparison with and without using a pre-processed parameter as input. On the right, the slope was involved in the clustering. On the left, we performed the clustering without using slope, then we added the existing slope to the clustered data to visually show the difference between using and not using slope.

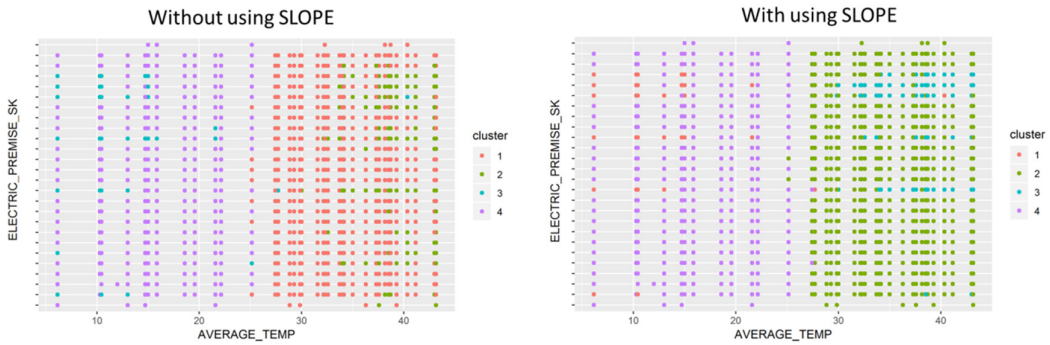

Figure 6 - The similarity of the results with and without using slope. Each row on the y-axis, Electric_premise_SK represents each customer. Only a limited number of customers presented here while the clustering was performed over 40,000 customers.

CONCLUSIONS AND FUTURE WORK

We have illustrated two approaches of using AMI data in distribution system load modeling. One approach uses AMI data accompanied by other household parameters to estimate the regression line for the dependence of energy usage on temperature and humidity. The slope of this line can then be used in clustering the customers. The other approach uses only the AMI consumption data and the temperature in the ZIP code to perform clustering. We showed that due to the fact that the consumption and temperature behavior of customers are already embedded in the slope, one might only use the data from customers to perform clustering and get almost the same results as using the slope.

Future work will focus on mapping the results to customers without installed AMI meters to see how accurate the prediction can be for those customers. Can a customer's monthly measurements (and thus load profile) be mapped to that of a similar customer for which interval data is available? Once this is known, the next step will be to infer the load factor for these customers based upon what is known about interval-metered customers with similar behavior.

REFERENCES

[1] https://www.ncdc.noaa.gov/sotc/synoptic/201801

[2] Hartigan, John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.1 (1979): 100-108.

[3] Wagstaff, Kiri, et al. "Constrained k-means clustering with background knowledge." *ICML*. Vol. 1. 2001.