Enhancing Customer Data Analytics by Leveraging Spatial Analysis and Third-party Data

Noel Stevens, DNV GL, Burlington, MA, Andy McCabe, AEP Ohio, Dublin, OH, Nathan Caron, DNV GL, Portland, ME, Richard Crowley, DNV GL, Portland, ME, Greg Gronski, DNV GL, Burlington, MA, Leon Roberts, DNV GL, Columbus, Ohio

ABSTRACT

This paper presents a novel approach to standardizing address information in customer profile, consumption history, and program tracking data to known physical locations, cleansing the account level data, and allowing data to be linked to third-party data with the increased accuracy. The challenge is that third-party information sources do not directly link to the utility's account list. The results of this process allowed us to construct a customer analytics database that defines each account in terms of their known consumption, participation, demographic, and physiographic characteristics. Furthermore, through leveraging the customer hierarchy information, we were able link individual accounts to distinct customers based on the location-parent company (parent-child) relationship provided by the third-party data. This robust data resource provides vital opportunities for utility companies to execute targeted marketing campaigns, direct face-to-face outreach programs, increase enrollment of non-participants, and reduce the cost of enrolling customers in energy efficiency programs.

We employ a multi-step process combining geographic information, phonetic algorithms, and logic rules to accurately match customer information available in both utility and third party databases. The matching process first uses geographic information systems (GIS) mapping software to map accounts to physical locations. To identify possible account pairs, we use geo-referenced numeric codes, phone numbers, and company and account names to cross-reference matches. The final step includes a decision algorithm that simultaneously checks each pair and picks the accurate matches between utility and third party databases. We use an integrated database to construct metrics such as energy and demand participation rates to assess the performance of customers within different segments as defined by customer attributes.

Introduction

When combined, utility consumption and program tracking data can provide a valuable resource for understanding customer behavior that can be leveraged to improve targeting customers for energy efficiency programs. When integrated with customer demographic and firmographic information, the utility data can be used for customer segmentation analysis, a powerful tool for identifying customers that represent opportunities for increased participation in energy efficiency programs or greater depth of savings of each participant. To leverage this information, however, the corresponding consumption, program tracking, and demographic data resources must be standardized, ensuring they report information for the same level of observation (i.e. premises, accounts, or customers) and that each record represents a unique observational unit in the respective datasets. Data standardization also helps improve match rates between utility and third-party data resources.

DNV GL was tasked by American Electric Power (AEP) Ohio, an electric utility serving portions of Ohio, to develop a new customer segmentation approach to improve their ability to target commercial and industrial (C&I) customers who represent opportunities for greater program participation and/or energy savings. In this paper, we discuss DNV GL's approach to standardizing information contained in AEP Ohio's separate consumption, program tracking, and third party data resources to improve match rates between the respective datasets and provide a robust set of data for customer segmentation analysis. We then employed a separate process that leverages various pieces of customer identifying information to link premise level records to distinct customers that represent centralized decision-makers or points of contact across sets of locations tied to the same business entity. Finally, we leverage these data to explore whether there are segments of customers who represent opportunities for increased participation or depth of savings.

Data Resources

In this section, we review the data employed in the segmentation analysis. Those data included three separate files received by DNV GL from AEP Ohio: 1) a customer information system (CIS) dataset; 2) an Experian dataset containing firm-o-graphic information; and 3) a monthly billing dataset. In addition, the analysis team received a fourth dataset of program tracking information from DNV GL's AEP Ohio Program Development and Implementation (PDI) group.

Customer Information System Dataset

The CIS dataset contained 185,318 observations with 149 variables that included the following fields: customer name, addresses, business types, tariff and tax information, and Standard Industrial Classification (SIC) and North American Industry Classification System (NAICS) codes.

Error! Reference source not found. provides a summary of the data standardization process of the CIS dataset. The standardization process started with all the observations and variables available in the original dataset. We checked if there are premises with multiple accounts and only retained variables that are important for the segmentation analysis.

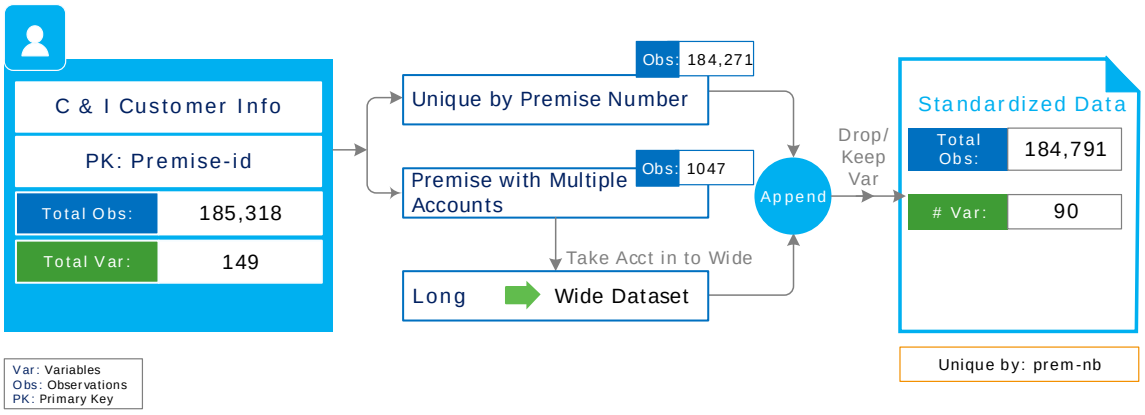

Figure 1. Overview of customer information system dataset.

Monthly Billing Dataset

The billing dataset contained monthly usage information along with bill start and end dates, revenue months, revenue class, and total bill and tax amount. The primary key identifier in the billing data was also the premise ID number. Figure 2 provides a summary of standardizing monthly billing data. There were 9,427,226 observations with 26 variables in the billing[1] dataset. The billing dataset is unique by premise ID number, tariff point, bill start date, and bill end date.

DNV GL found that there were 1,551 duplicate readings when sorted by premise number, tariff point, bill start, and bill end date. These duplicate records were removed from the analysis dataset. The final monthly billing dataset contained 9,425,675 observations and 16 variables.

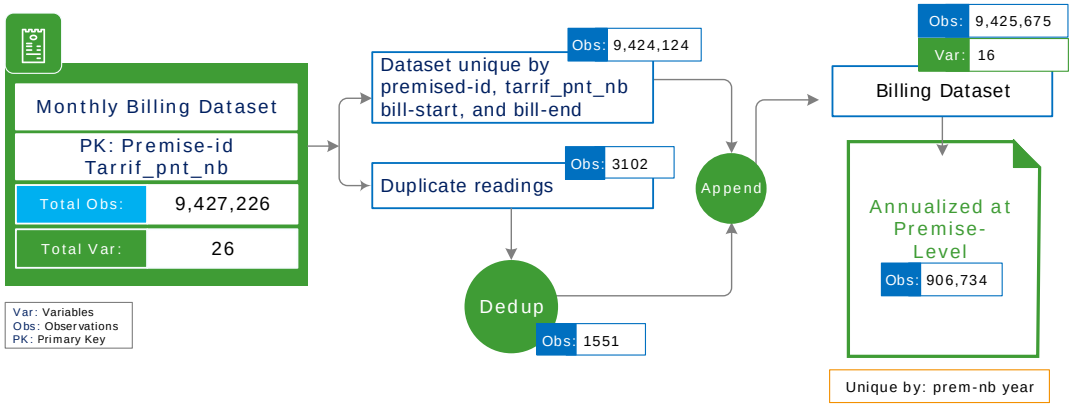

Figure 2. Overview of monthly billing dataset.

Experian Dataset

There were 116,085 observations (unique premises) and 79 variables in the Experian dataset, which contains the following four sets of information:

- Customer-identifying information such as business and contact names, service addresses, phone numbers, premise ID number, parent-child relationship information;
- Business-related information such as NAICS and SIC codes, the number of employees at the location, annual sales, and business start year;
- Credit related information, including bankruptcy status, derogatory indicator, and liability amount; and
- Geographic info, including addresses, latitude, and longitude.

Figure 3 shows the standardization process of the Experian dataset. The information in the Experian dataset is unique by premise number. DNV GL looked at each variable available in this dataset and decided to keep 49 variables for the segmentation analysis.

[1] The name of the raw monthly billing dataset is monthmerge.

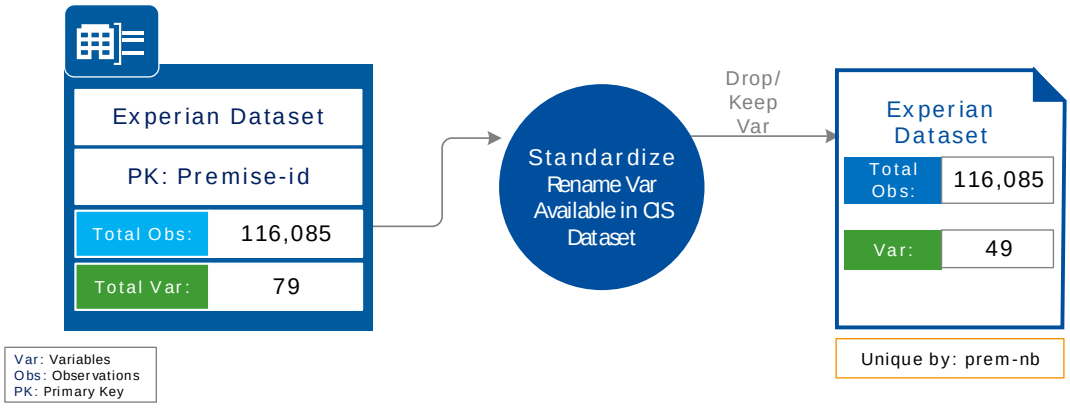

Figure 3. Overview of Experian dataset.

Tracking Dataset

Tracking data contains customer information, project status, program types, project cost, and energy savings maintained by DNV GL's PDI group. The PDI group maintains AEP Ohio's tracking dataset both at the project-level and measure-level. We primarily use project-level tracking data in the segmentation study.

Error! Reference source not found. shows the standardization process for participation data. DNV GL appended different measure-level information in the project-level tracking dataset. The final, standardized tracking dataset is at the project-level and has 17,715 observations with 135 variables.

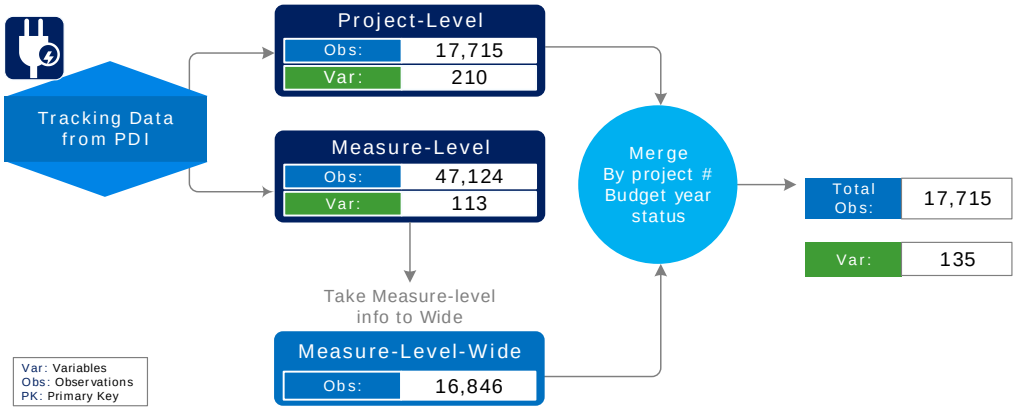

Figure 4. Overview of program tracking dataset

Data Cleansing and Account Matching

This section provides an overview of the data cleansing and account matching process between AEP Ohio's CIS and business related Experian data. The goal of the matching process was to try to match CIS premises Experian had not matched to the records they provided in the Experian.

DNV GL performed matching using multiple variables such as addresses, names, and phone numbers available in both datasets. We then devised a decision algorithm to pick accurate matches based on criteria developed by DNV GL.

Overview of Data Cleansing and Account Matching Process

Figure 5 provides a high-level summary of the process used to match additional premises from the CIS database with the Experian dataset. The process starts with matching addresses available in both datasets with the standard addresses of US Census TIGER files, which creates geocoded reference IDs in the matched addresses. DNV GL employed software that links records in the CIS and Experian datasets based on similarities in address fields such as addresses and account names. Then, DNV GL performed logic-based matching based on addresses, names, and phone numbers to pick the accurate matches available in both the Experian and CIS datasets.

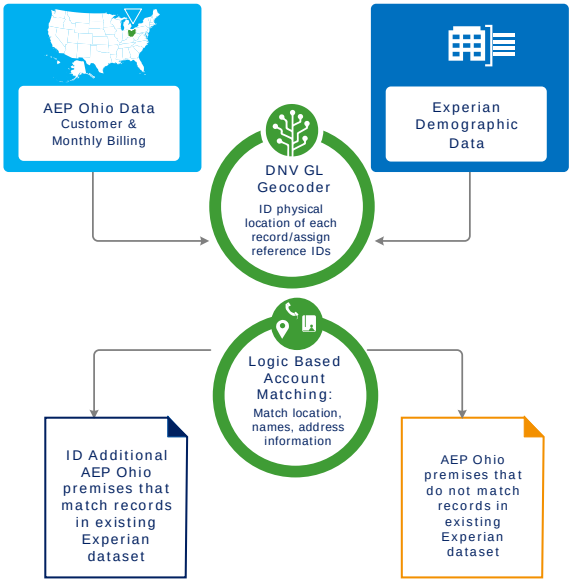

Figure 5. Overview of account matching process.

Geocoding Process – Identify Physical Locations of Records

Geocoding is a powerful tool that we used to standardize addresses based on the information from the US Census TIGER files, and then provides accurate coordinates that place the premises to a digital street map of an approximation of the precise geographic location of a specific address. We used geocoding to see how many of the addresses from each dataset matched a physical location on the map. The geocoding process allows the user to accept address matches at different levels of accuracy, depending on the level of precision required. For this study, we examined the geocoded results at 90% and 100% accuracies. This level of accuracy allows minor spelling differences in street names, but does not accept addresses where zip codes match but city names differ or city names match but zip codes differ. The geocoding process creates reference IDs (REF_ID's) that are unique identifiers to show if an address is matched to a TIGER point in both the CIS and Experian datasets.

Table 1 shows the results of creating reference IDs in the geocoding step between the two datasets and the TIGER files based on their spelling sensitivity setting.

Table 1: Geocoding Status of addresses available in Customer and Experian dataset

Description	CIS Dataset	Experian Dataset
Total Premise Count in CIS dataset	185,318	116,085
Percentage of Premises that have geocoded reference IDs at 100% Accuracy	69.8%	68.4%
Percentage of Premises that have geocoded reference IDs at 90% Accuracy	81.5%	79.8%
Address Variable	Serv Street_Ad	serv_street_ad
Zip code variable	Zip5	zip5

We also matched addresses of the CIS and Experian datasets with the help of logical matching functions in SAS. The logical matching process between the addresses available in the CIS dataset and Experian dataset help capture addresses that may be in different formats or spellings. An example of this would be a misspelling of main street ("main street" and "mian street" in the same town are likely the same place, but a data entry error shows them as two different places). DNV GL employed the logic rules[2] to pair addresses within the same zip code that are similar. We then generated match scores[3] using paired addresses to see how closely the addresses match between the two datasets. The score gives the degree of difference between two addresses, computed by comparing character by character of paired addresses. Since the score is also a function of character length of the addresses, we normalize the match scores by dividing by the length of the CIS address. The match score of zero indicates a perfect match; as the magnitude of the score increases the accuracy of match decreases.

Similarly, we calculated match scores between account names available in both datasets. We used three different names from Experian dataset – customer name, business name, and executive name – with the account name from the CIS dataset to calculate the match scores. DNV GL also matched the two datasets using phone numbers and zip codes. We then utilized the matching dataset created by matching in four different ways (address, geocoded reference IDs, phone number + zip code, and account names) to perform checks and pick accurate matched pairs.

Checking accuracy of matches

DNV GL devised a decision algorithm by simultaneously checking four different matches for each paired address with the goal of choosing only accurate matches. Error! Reference source not found. contains the detailed flowchart regarding the decisions of picking the accurate matching pairs between CIS and Experian dataset. The initial matches between CIS and Experian dataset are based on the following criteria:

1) Geocoded reference IDs obtained from ArcGIS

[2] Roesch, Amanda. Matching Data Using Sounds-like Operators and SAS Compare Function. SAS Global Forum 2012. http://support.sas.com/resources/papers/proceedings12/122-2012.pdf

[3] Staum and Waldron. Fuzzy Matching using the COMPGED Function. NESUF 2007. http://www.lexjansen.com/nesug/nesug07/ap/ap23.pdf

2) Logical matching on addresses using sound-like function in SAS using standard data matching commands that also provide levels of precision to matches
3) Phone numbers and zip codes
4) Account names

The geocoded reference IDs and phone numbers are numeric values, thus making it easier to check the accuracy of matches. We used the match scores that are below a certain threshold to pick an accurate match based on paired addresses and account names. To define the threshold limit for comparisons that are deemed matches, we compared a sample of the observations from the CIS and Experian datasets at different score values and found a match threshold that gives us accurate matches.

Matches based on a single piece of information do not always give the most accurate matching. For example, we found instances where geocoded reference IDs were the same for two different premises that have the same street names, but different address numbers. Similarly, we also found a large business or city office reporting the same phone numbers for their multiple premises at different locations. Our decision algorithm is based on choosing accurate pairs by looking at a minimum two different matching criteria. More specifically, a pair is considered accurately matched if:

- Geocoded reference ID match and name match score is below certain threshold; or
- Address match score and name match score are below the threshold value; or
- Geocoded reference ID or Phone numbers match and Address match score is below the threshold value.

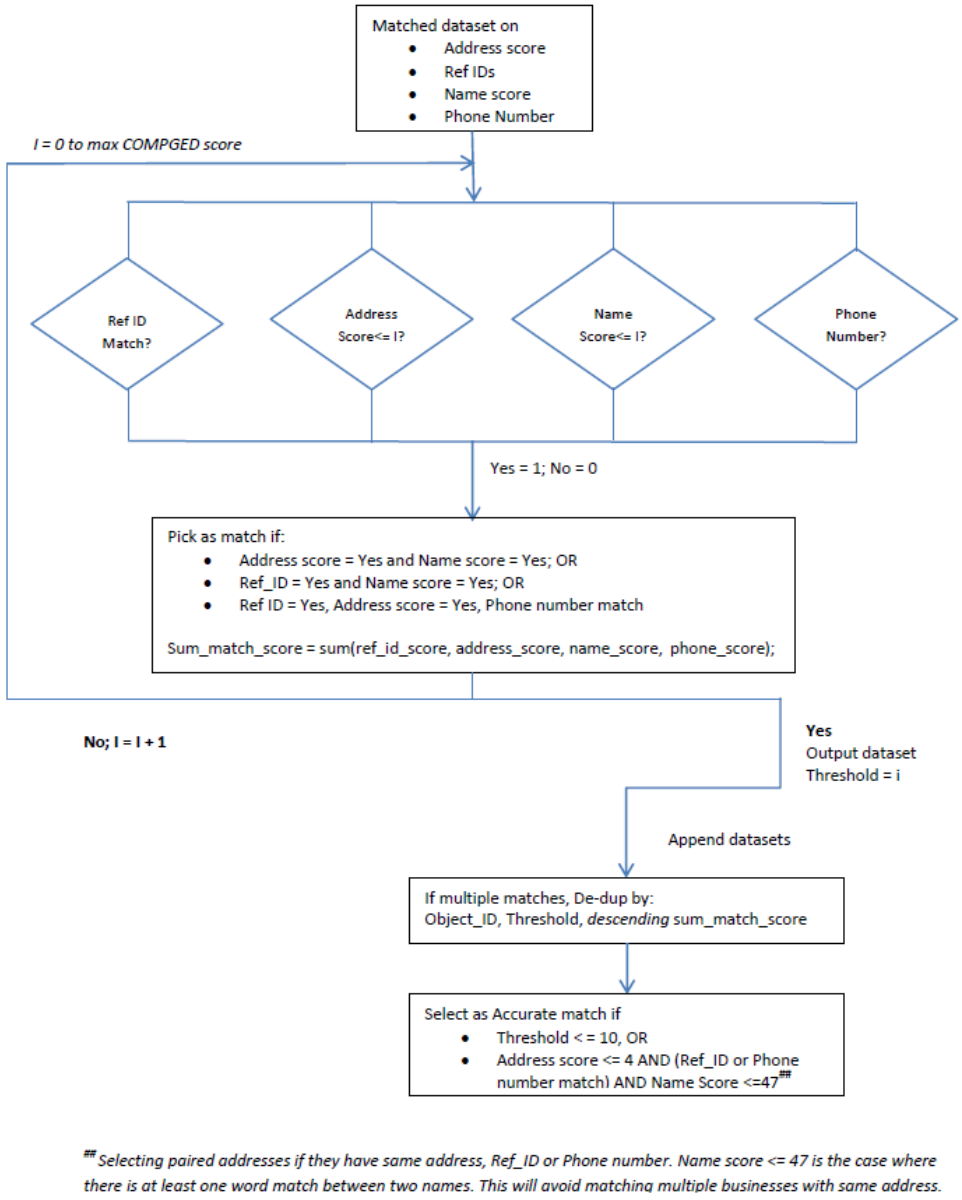

Figure 6: Logical decision for choosing accurate matching pairs.

Account to customer level

Identifying a centralized decision maker across accounts or premises can improve customer management and possibly lead to greater success in marketing energy efficiency programs. Historically, utilities have viewed their customers in terms of the boxes on the sides of buildings (accounts or meters). However, these boxes are not "customers," and C&I customers do not operate in order to consume electricity. Rather, C&I customers operate for a business purpose, such as to sell groceries or manufacture cars. These "customers" will often have multiple accounts that may be

located within the same building or at different addresses throughout the utility's territory. Examples of these complicated situations include the following:

- Within a single building, such as an office park, campus, hospital, or mall, the same customer may have multiple offices or locations that each have separate accounts.
- Within a utility territory the same customer may link to multiple accounts at separate locations, or even within multiple floors of the same location.

The ability to link multiple accounts into distinct customers allows a utility to examine the full consumption, program participation history, and demographic profile across the entire business. This enables the utility to better understand the corporate decision making process regarding energy efficiency and identify the key decision makers at the centralized headquarters or franchise owners. DNV GL leveraged data from both the Experian dataset and the AEP Ohio customer dataset to tie individual accounts to distinct customers. We used the following logic to define two separate customer IDs depending upon which data were available for each record:

Customer_ID1:

- Where there was an Ohio based parent ID in the Experian data, then Customer ID1 = Experian Parent ID
- Where no parent ID was available in the Experian data, then Customer ID1 = AEP Ohio's customer number

Customer_ID2:

- Where Experian had assigned a parent ID and that parent was in Ohio, Customer_ID2 = Parent ID + telephone number. This was to restrict customers to a level that had the same contact information in the Experian data, as a sublevel to the overarching corporate parent
- Where Experian had NOT assigned a parent ID then:
    - If Experian assigned a company number, then Customer_ID2 = Company number + telephone number
    - If Experian had not assigned a company number, then assign based on telephone number only
    - If Experian had not assigned a Parent ID or a Company number and there was no phone number, then assign based on AEP Ohio's customer number

Finally, we compared each of the IDs for each record to identify instances in which the IDs did not agree. In these cases, we select the preferred ID based on a set of logical rules that set priorities based on the matching fields for each record.

Results

Data Cleaning / Account Matching Results

Both the CIS and Experian raw datasets that DNV GL received contained prem_nb as a unique identifier. This variable can be used to match between CIS with Experian dataset. When these two datasets were matched by using prem_nb, only 60% (111,104) of the premises available in the customer information data were matched with the Experian data. The remaining 40% (74,214) of the premises will not have information from Experian data. Our matching process could pair an additional 2,616 premises (out of 74,214) from CIS dataset with the Experian dataset that otherwise would not have a match.

DNV GL found 71,598 premises do not have a match with the Experian dataset provided. Out of 71,598 premises, DNV GL could get geocoded reference IDs for 68.5% of these premises with accuracy set at 100%. The percentage of premises with geocoded IDs increases to 80.2% when the spelling accuracy was reduced to 90%.

DNV GL provided AEP Ohio with all records that Experian was not able to find corresponding demographic information for, but that we were able to identify through the Census' TIGER file. Experian used this information to identify roughly 2,400 additional records with matching demographic information.

Customer-Level Analysis Results

The shows the impact of moving from an account (or premise) to a customer-level analysis. Figure 7 compares the number of premises with different consumption levels to customers whose aggregate consumption across all linked premises fall within each size category. The data show that the customer level view of the data provides a greater percent of large customers than the premise level view. This finding has implications for how companies like AEP Ohio can target these customers through a centralized decision maker, thereby providing improved customer management and possibly greater success in marketing energy efficiency programs.

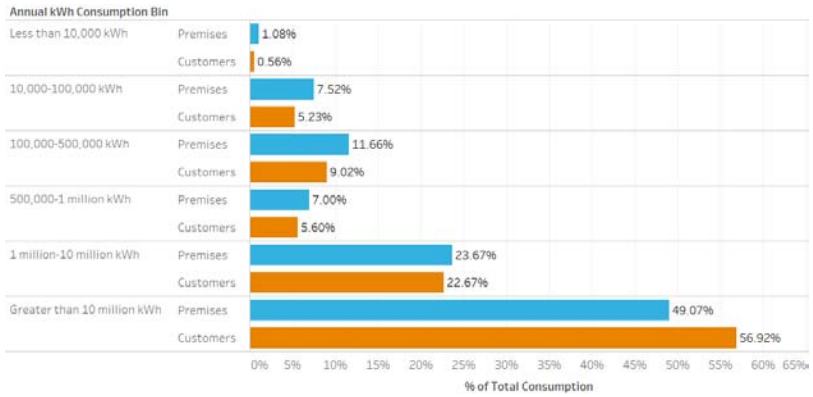

Figure 7. Distribution of overall consumption by consumption bin: Comparison of customer and premise level data.

Figure 8 presents a geographical display of the savings ratio of customers by zip code, where the savings ratio is defined as (Total kWh savings from participants) / (Total consumption of participants). This information can be combined with other data such as measures installed by different participating service providers, account representative sales figures, or demographic information to provide valuable insights into where resources are being allocated effectively, and where additional resources are required. For example, if internal sales or external service providers are geographically aligned, then a graphical display of the data can help call out regions where additional resources may be required. Zeroing in on specific counties, towns or block groups can assist in isolating populations that may provide opportunities for increased savings by existing participants.

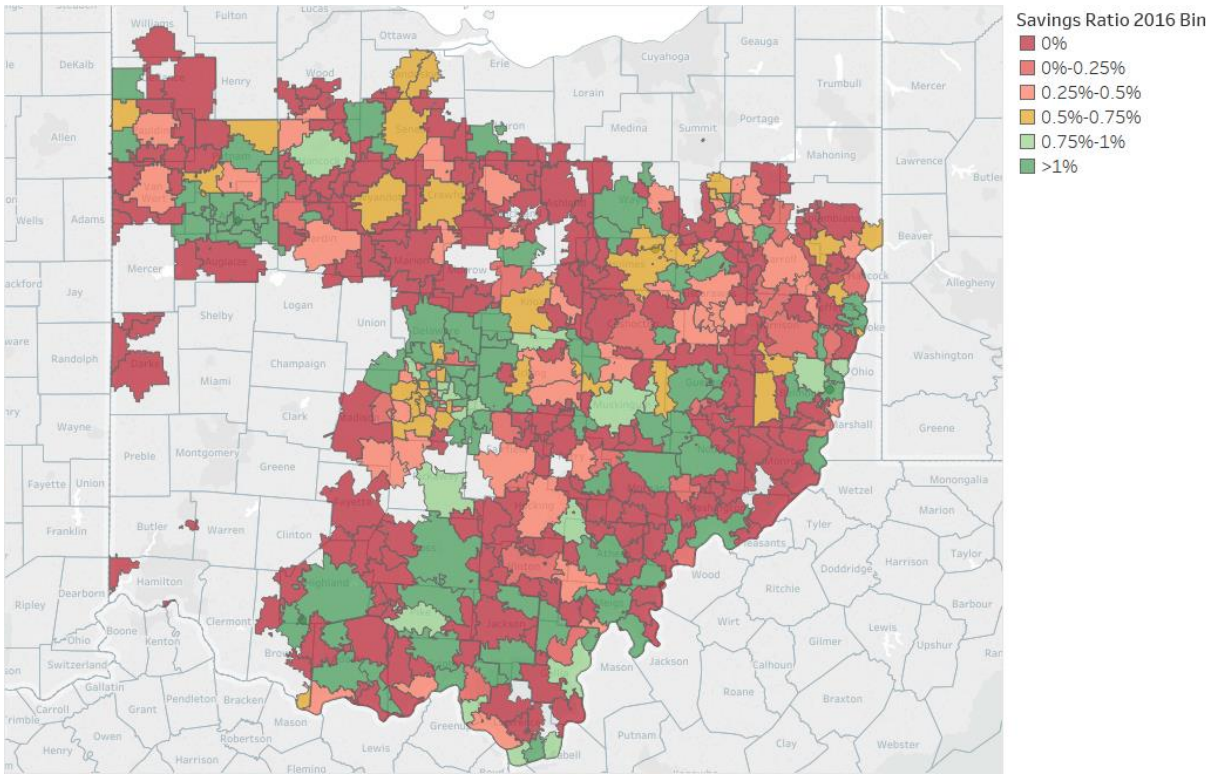

Figure 8. Distribution of savings ratio (participant savings / participant consumption) by zip code

Industry Segmentation Results

Figure 9 presents an example of the improved customer analytics capability that leverages the third party firmographic data. The figure contrasts the participation rate (y-axis) verses savings ratios (x-axis), and number of premises in the customer segment (bubble size) within various sub-industries of the manufacturing sector.

The data show that customers in the machinery and equipment industry have a relatively high savings ratio but low savings weighted participation rate. Customers in this sector represent opportunities to increase participation, as each participant provides substantial savings potential. In contrast, customers in the non-ferrous metals, petroleum products, and other fabricated metal

industries have relatively low savings ratios but high energy weighted participation rate. Firms in the resin and synthetic rubber sector have both a high savings ratio and relatively high energy weighted participation rate, indicating that these customers offer limited opportunity for increased participation or deeper savings.

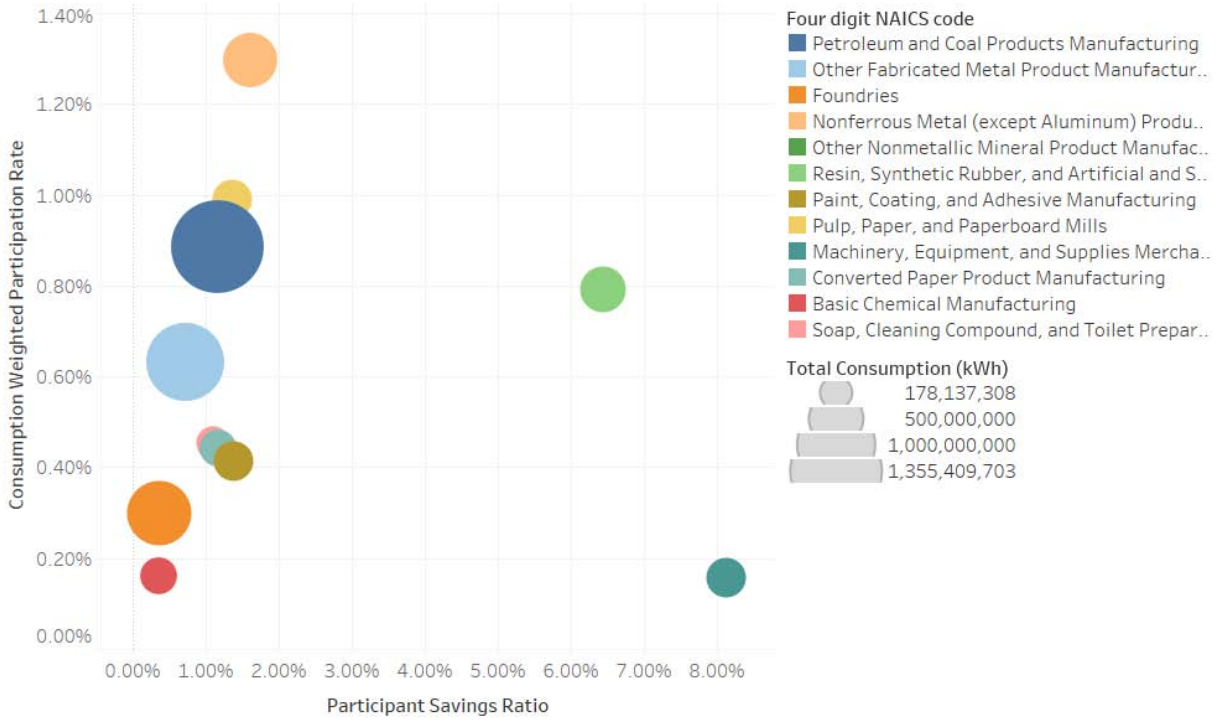

Figure 9. Savings ratio versus participation rate by sub-industry: Manufacturing sector

Summary

Customer analytics can provide program designers and implementers with valuable tools for identifying untapped savings opportunities. In this paper, we presented DNV GL's approach for leveraging special data analysis and logical data processing to increase match rates between premise level consumption and program tacking data, and demographic profile data. We then leveraged information provided by Experian to provide a customer level view that combines multiple premises tied to the same customer, or decision maker. Finally, we used data analytics to isolate groups of customers who may represent opportunities for increased savings.

The ability to link consumption, program tracking information, and third-party demographic information can greatly enhance efforts to profile customers who represent opportunity for increased participation or deeper savings. Further, our approach can be used to improve the accuracy of geographic analysis in general, as CIS accounts without locational information will not be included in mapping analysis. This is particularly an issue if non-matched has systematic causes. Having accurate addresses has benefits that extend well beyond energy saving. Time spent by repair folks to find locations, support in fire and other emergencies.

References

AEP Ohio Market Segmentation Final Report. Prepared for AEP Ohio. Prepared by DNV GL. May, 2017.