Memo to:
Massachusetts Program Adminstrators Research Team and Energy Efficiency Advisory Council EM&V Consultants

Prepared by: Rich Crowley, DNV GL

Date: Original - December 28, 2018
Revised – February 7, 2019

Enhanced Customer-Level Database Capabilities

Summary of project activities and database use cases

1 INTRODUCTION – PROJECT BACKGROUND

DNV GL maintains the Massachusetts Commercial and Industrial (C&I) Evaluation Database for the Massachusetts Program Administrators (PAs). The MA C&I Evaluation Database captures the annual billing, customer, and installed energy efficiency measure information by program year, including energy consumption, savings, geographic data, and contact information at the account level, to support evaluation, measurement, and verification (EM&V) projects in Massachusetts. DNV GL uses these data to generate the annual MA C&I Customer Profile Study report, a roughly 400-page document that breaks out past years' energy efficiency activities into detailed tables, charts, and maps.

In the 2015 C&I Customer Profile Study reporting cycle, the PAs and Energy Efficiency Advisory Council (EEAC) Consultants approved a pilot effort to collect a limited series of tax data to create energy use intensity tables for different building types at the location level as part of a new analysis segment in the C&I Customer Profile report. The pilot's success, coupled with insights gained through the pilot data, led to the PAs' Enhanced Customer-Level Database Capabilities project (project). This project carried out a larger data integration effort that captured data at various levels (account, location, customer, cross-fuel), increasing the value of the MA C&I Evaluation Database for EM&V work.

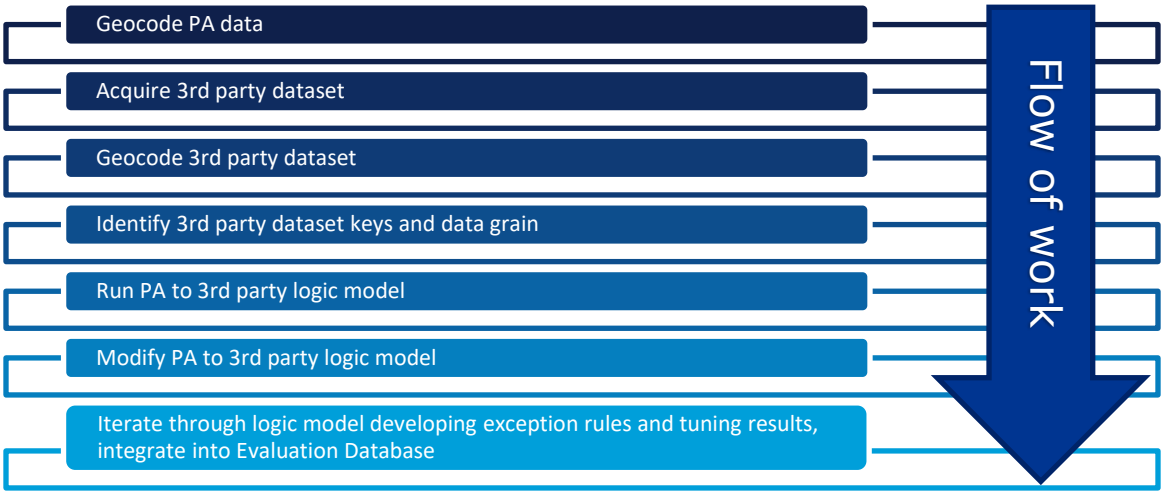
Although the Enhanced Customer-Level Database Capabilities project developed the C&I Evaluation Database, it could be also leveraged to improve the MA Residential Customer Evaluation Database that DNV GL maintains as well. Accordingly, the newly integrated data was made available to DNV GL's Residential team, and has been used in a number of Residential and C&I projects.

2 PROJECT DESCRIPTION

The project identified and acquired third-party data to support ongoing EM&V efforts. It also developed the iterative logic model that uses geographic and string-search algorithms to perform

increasingly accurate levels of matching between datasets.[1] The logic model's flow of work is detailed below in Figure 1, followed by a matrix in Figure 2 detailing third-party data sources, brief descriptions of their contents and the contract area (Residential or C&I) that could benefit from them.

Figure 1. Third-party data source work flow



[1] It is necessary to match datasets using similar identifying variables when the datasets do not share a common key. This was the case with the PA-provided data and third-part datasets. While these datasets did not share a key, they did contain variables such as customer name, address, and phone number that could allow for matching. However, since these variables don't come from a shared data source, they can contain slightly different or incorrect information. For example, the address 10 Main Street and 10 Main Street East might be the same location named differently in two datasets. While a direct merge would not result in a match, our matching algorithms likely would. Similarly, the names John Smith, John Smiht, and John Q. Smith located at the same address likely refer to the same individual, but would also not result in a match based on a direct merge (unless the matching algorithm accounted for spelling differences or omitted information in matching).

Figure 2. Matrix of third-party data sources, their contents, and contract areas that could benefit from the data

Data source	Contract area	Brief description
Emergency 911 Data	Both	Detailed point-level data of all building and sub-locations in Massachusetts; used as a critical input into the accurate geographic coding of address in the PA and third-party data sources so the logic models could leverage proximity as part of the scoring assessment.
US Census TIGER Data	Both	Detailed range-level geographic data served as a back-up input for the accurate geographic coding of address in the PA and third-party data sources when the e911 data could not identify an acceptable match so the logic models could leverage proximity as part of the scoring assessment.
MA Level 3 Tax Assessor Data	Both	Detailed geographic and supporting data attributes for all recently assessed tax parcels in the state. Includes building characteristics (age, style, size, type) used in the data mining, reporting, and dashboard products for the PAs. Also includes owner information used to improve the logic models; matches rates to the utility data records and provide a proxy for shared ownership buildings.
InfoUSA Data	C&I	Local geographic data and supporting attributes for many of the identified businesses operating in the state. Includes building information (industry sector, employee and sales sizes, credit score) used in the data mining and reporting for the PAs. Also includes owner and parent corporation information used to improve the logic models and match rates to the utility data records, provide a proxy for commercial real estate buildings, and identify PA accounts and locations within national chains and franchises.
Dodge Players Database	C&I	Identified the key businesses and types of new construction or significant renovations occurring; augmented the tax data by providing a more recent dataset to reflect changes in building stock and size.
EPA Facility Registration System	C&I	Environmental interest variables including emissions level, fuels, and compliance. The FRS data exists as its own larger, separate database environment with industry variables, contact information, program information, and other data useful for targeting large industrial locations. DNV GL integrated the keys from this data environment into the PAs' C&I Evaluation Database for seamless communication between datasets.
Zip Code Business Patterns Data	C&I	Zip code business patterns allow for aggregate and regional comparisons of account groupings by industry sector and employee size bin. Useful in understanding long-term industry trends and in modeling efforts. PAs also use this data to cross-validate industry feedback on long-term market trends that can affect willingness to invest in energy efficiency measures.
American Community Survey Data	Both	ACS data provides numerous local aggregated demographic variables that have been useful inputs for geographic data mining and targeting. This material has also helped identify accounts that are likely to be delivered fuels (though a combination of probabilities from different datasets) and understand low to moderate income populations in the PAs' service territory.
Energy Star® Buildings Data	C&I	Provides a year, style, and square footage as well as an Energy Star score for the Energy Star-certified buildings in MA. DNV GL integrated with the PA Evaluation Data and tax data to facilitate the ability to generate an approximation of a benchmark score for the different tax locations in MA.
LEED® Buildings Data	Both	Provides information on the certification level and status of LEED buildings in MA. In conjunction with the tax entity data, this dataset can help inform what buildings or owners may have a positive outlook on undertaking energy efficiency or environmental measures. Through the customer, key integration can be used to identify what other buildings in the state they may have in their portfolio.
Cross-PA Data	Both	This information helps links accounts and locations across PAs, sectors, and fuels for a more comprehensive picture of how energy is used and how measures are undertaken over time. Although not a third-party dataset, the PA evaluation data was the foundational target that all third-party data was integrated against, and is a critical component of the integration project.

3 PROJECT OUTCOMES

This project developed a series of logic models that integrated third-party datasets with the PAs' C&I and Residential customer information systems, to help PAs and evaluation contractors better understand the customer population. Integrating third-party data helped develop additional analysis grains including location, tax parcel, and cross-fuel customers. It also provided a large, documented, and standardized set of new variables that evaluation projects could incorporate into data mining and analytical undertakings.[2]

Prior to the data integration, DNV GL conducted a series of in-depth interviews with PA staff to better understand key data needs and overarching benefits the PAs hoped to achieve via the database enhancement effort. We identified 9 overarching benefits (Table 3-1), most of them realized in the 2018 follow-on projects.

Table 3-1. PA-identified key data needs and benefits from the database enhancements

Benefit
Improve normalization of customer-level energy consumption
Improve service and program offerings to multi-site customers
Provide cross-referencing for customer matching while preserving customer confidentiality
Leverage economies of scale by linking customers, integrating data and classifying customer attributes at a single point in the process
Provide all PAs with a standardized set of firm-o-graphic attributes
Enable prospect modeling capability for non-participating customers leveraging cross-PA participation data
Identify opportunities for deeper measure-level penetration using participation information from other PA territories
Enhance survey and site visit value for campus locations
Facilitate two-way flow of data between DNV GL and PA teams

One of the most impactful outcomes for this project was the integration of the 351 MA towns' tax assessor data. The project significantly increased the evaluation contractors' integration rates, while decreasing the update cycle length from months to days or weeks. The driver of this reduction is that the data integration process DNV GL developed can be reapplied to updated version of the data, often with only minor modifications to the code. This has the added benefit of not only providing more timely access and insights using the data, but also allowing future investments to focus efforts on fine-tuning the output results for enhanced data accuracy.

Through the project, DNV GL was also able to more effectively integrate cross-PA account, customer, and site-linked data. With PA guidance through working groups, this project developed a process for identifying which data variables, and at what aggregation level, material could be coordinated across PAs. This cross-coordination helped increase customer insights and improve customer experiences.

A striking example of helpful cross-coordination involves understanding site-level participation across electric and gas PAs. Developing the data sharing framework gave the PAs better insights into which customers received utility-served electric and gas (rather than delivered fuels), and into whether gas customers had

[2] Prior to the database integration project individual evaluations that wanted to leverage third-party datasets had to identify the datasets, develop their own integration logic, and proactively verify that data was re-integrated into the Evaluation Database. By centralizing the integration under a single dedicated project the PAs reduced the potential for duplicative efforts across projects and increased the cost effectiveness of both acquiring and integrating individual data sets. As more projects continue to leverage this information, the value added though this centralization will continue to increase.

previously participated in energy efficiency programs under their electric PA, or vice versa. This allowed for more strategic targeting of past participants and more informed designs for future energy efficiency efforts.

Another lasting outcome of the cross-coordination will be better integration of customer contact flags for survey work. For instance, the logic models and cross-fuel linking could allow customers who have asked their gas PA not to contact them to have that preference shared with their electric PA as well. Such coordination can potentially increase customer satisfaction while also improving evaluation studies though greater population screening.

During 2018, the PA and evaluation teams began using the newly integrated data for several of the more advanced future analyses envisioned in the customer-level value proposition and third-party data integration value propositions from the original work plan. The most successful example of this has been one PA using the integrated data, augmented by its own internal customer information, to identify customer groups who represent opportunities for targeted weatherization program outreach. The efficiencies gained by having data standardized and ready for analysis allowed this effort to be undertaken in a series of monthly sprints, with results quickly available to inform implementation, strategy, and management teams. Preliminary results of this effort include a follow-on study seeking to better understand barriers for certain customer groups, and to update the analysis from the residential population to explore C&I customers with delivered fuels.

Section 4.1 presents a selection of projects that DNV GL conducted with the PAs, using data in ways that were not possible prior to building the Enhanced Customer-Level Database. While we did not conduct a formal process to quantify the uses and value of the Enhanced Customer-Level Database to these projects, we have received anecdotal feedback from the PAs characterizing the data's various uses and value.

4 USE CASES AND STATUS

4.1 Projects leveraging the Enhanced Customer-Level Database

The following sections present a series of projects that took advantage of various datasets integrated through the Enhanced Customer-Level Database. These projects illustrate the different ways that the PAs and the DNV GL team have used the data in both the C&I and Residential contracts, and can be mapped back many of the overarching benefits articulated earlier in Table 3-1. They also show how PA strategy and implementation teams have leveraged the database investment outside of the EM&V space.

4.1.1 Information response support for time-series economic trends in PA service territory

One of the gas PAs noted anecdotally that economic conditions in its service territory were causing its C&I customers to reduce their capital investments. This in turn affected C&I customers' willingness to invest in larger energy efficiency infrastructure, such as HVAC and process measures. Regulatory stakeholders asked this PA to provide supporting macroeconomic data on territory-wide time-series trends in employment and sales by industry sector.

The impacted PA's team lead was familiar with the data integration project, and knew that zip code economic data, including industry, employment, and sales volume data, had been integrated into the data ecosystem. The PA lead engaged DNV GL to quickly and cost-effectively respond to the information request

by populating a template with the data. The PA lead coordinated with DNV GL to extract a detailed industry sector table at the zip code level over a 10-year period, and to provide a comparative snapshot of the same industry sectors across the other PA territories. This satisfied the information request a few business hours after the PA had received it.

4.1.2 Building-level energy use intensity mapping

Early on, DNV GL used the project to identify a cost-effective way to provide a normalized view of energy consumption data across the PA customer populations. DNV GL assessed numerous third-party data sources with attributes such as building square footage and building vintage, and determined that the best option was a combination of the MA Level 3 statewide tax assessor data and the city of Boston's tax assessor data. The determination was based on these data sources' detailed statewide coverage and their use of MA-specific standardized values. These sources are free to use, and they represent no maintenance or data collection burden on the PAs.[3]

For MA locations, DNV GL uses tax parcel location keys and square footage for all tax records to provide both "within fuel" and "combined gas and electric (MMBtu)" views of the locations' average energy consumption per square foot. This allows the PAs and stakeholders to identify localized sub-populations where the energy use intensities (EUIs) are consistently higher than comparison areas, and to zero in on the specific buildings and accounts that drive these geographic hot spots. The location-level keys also offer a transparent way to link gas and electric accounts without shared key values across the PAs, providing a more comprehensive view of locational energy consumption rather than focusing on a specific fuel. These insights can help PA teams understand whether an account is a potential candidate for measures like heat pumps, or whether it represents a location whose previous occupant installed an efficient gas furnace and may therefore not be a good candidate for heat pump offerings. The location-level EUI maps are part of the annual Residential and C&I Customer Profile Study reports, and PA teams frequently use the detailed underlying account information in evaluation and implementation efforts.

4.1.3 Identification of state-wide building types

The PAs use a mix of classification systems to describe accounts within their populations. These include NAICS codes, SIC codes, internal alpha-numeric codes, and other third-party classification systems. Historically, DNV GL used a matrix of the different classification systems to arrive at a "best" representative classification for statewide and cross-PA comparisons based on the 21 standard NAICS codes in the C&I population. The integration of the L3 tax assessor data gave the PAs a series of MA-specific detailed building use codes that ensured, for example, that what a tax assessor classified as "large supermarket" in a small gas PA's service territory was consistent with what was classified as a "large supermarket" in the Boston area's electric market. It also gave the PAs a continuously updated dataset reflecting changes in their populations; the codes developed allowed updates to take only days rather than weeks or months.[4] The identification of statewide building types has supported multiple use cases, detailed in the following sub-sections. Residential and C&I evaluation contractors have also used this data as part of an interactive PA multifamily cross-sector tool.

[3] While the data is free to acquire it does not natively have keys that link it to the PA customer data. This project developed and refined a series of detailed logic models that integrated this data with a high degree accuracy and a minimal number of false positives. These models and codes are what is reapplied to the tax data though the update cycles.

4 Since the start of the project, the tax data has gone through two additional update and integration cycles using the existing code and framework DNV GL developed, most recently in September 2018, to ensure that teams are accessing the most recent tax data and integrating this with the PAs data.

4.1.4 Evaluation surveys and population weighting

Having a consistent classification system of statewide buildings improves the ability to draw detailed samples and weighting for surveys. It also allows residential populations to be divided into classes that did not generally exist in the data available to the PA energy efficiency teams. A recent example of how these codes have been used for sample design is in the residential baseline survey, where DNV GL was able to provide the PAs' evaluation contractor team with a statewide population of standardized building use codes from the integrated MA tax assessor data; the team used these codes to improve the representative survey population by providing more granularity on building types (small vs. large multifamily) and greater consistency across the state population through the use of a standardized building code. The codes have also complemented the completed C&I baseline study.

4.1.5 Cross-PA measure targeting

The consistent classification also helps the PAs more effectively ask and answer questions like, "What types of energy efficiency measures did this type of customer install in different service areas?" DNV GL has been able to support PA implementation teams seeking to understand what measures specific customer types have installed (e.g., "Is any customer type installing wi-fi thermostats more frequently than the population mixture would predict?") and build on this by answering the forward-looking question, "Who are the customers in this building type that have not installed these measure yet?"

4.1.6 Identifying multifamily households for PA dashboard

Multifamily households can be served though a specific multifamily PA program offering. Integrating standardized tax use codes has allowed the PAs to identify the location of multifamily buildings with more certainty. Beyond the binary identification, the use codes have also facilitated a better understanding of the type and style of multifamily buildings that exist in the PA population (for example, distinguishing between condominium units versus duplex or triple-decker homes). PA evaluation contractors have used this level of detail to develop a geographic multifamily dashboard for the PAs to help interpret the results of a recent multifamily evaluation.

4.1.7 Data mining inputs to PA weatherization customer targeting

Two PAs have asked DNV GL to use integrated datasets to better understand who has participated in their weatherization offering. One PA sought to understand various barriers customers have encountered in order to modify program design and customer targeting to improve weatherization audit-to-closure rates. The PAs used a combination of historical PA billing and tracking data, implementation audit and barrier data, tax data, and customer physiographic data[5] to develop a blended prioritized list of weatherization outreach targets. One PA elaborated on this model by developing a series of geographic insights and linked data across electric and gas accounts to beta-test the targeting of dual-fuel accounts. Results of this data mining effort were presented in New Orleans at the 2018 AESP National Conference.

[5] Although one PA provided this data to use with the data matching models developed in this project, it is not implemented into the larger data ecosystem due to contractual restriction on this purchased data.

4.1.8 Data mining inputs to PA delivered fuels customer targeting

Integrating cross-fuel linked data, ACS data, and building characteristics data helped DNV GL and the PAs identify accounts and locations that are likely to be served by delivered fuels.[6] These customers may represent higher priorities for things like energy optimization outreach.

DNV GL developed a model to determine a location's total energy consumption and weather sensitivity for electricity (not gas) usage. The goal was to identify accounts whose electric consumption was highly correlated with temperature, and who would thus be likely candidates for electric resistance to ASHP energy optimization.

We then identified whether a building used a delivered fuel based on rate codes and whether an area was served by natural gas. We refined the model using additional third-party data, and applied the refined model to the statewide population of one PA's customers in all duel-fuel towns. This work will be used by future Customer Profile reports, particularly on the residential side.

4.1.9 Identifying national accounts

Cross-fuel linked data, together with corporate parent identification numbers from the linked InfoUSA data, allowed the PAs and DNV GL to identify large national accounts and other locations associated with them, across fuels and across PAs. These accounts may have decision-making procedures (e.g., in corporate purchasing policies or centralized purchasing departments) that do not take full advantage of MA energy efficiency offerings. These accounts may also represent priority opportunities for multi-year energy efficiency engagements though mechanisms like memorandum of understanding with the PAs.

This material was recently used to identify MA customers that were part of larger corporate entities, in a study evaluating how MA customers performed relative to their national peers. Prior to the integration of the InfoUSA data and cross-fuel linked data, such an effort was significantly more time-consuming, and did not have the same comprehensive population coverage.

4.2 Potential future uses

The following sub-sections detail some examples of potential future uses for the Enhanced Customer-Level Database based on PA feedback, stakeholder comments, and industry trends. While these sub-sections discuss general tasks and examine why and how the database could be useful for these tasks, they do not lay out any methodology, as we have not scoped these tasks outside of working groups and brainstorming sessions with PAs and EEAC Consultants.

4.2.1 Customer segmentation and classification

The large volume of data aggregated though this project substantially improved the ability to retroactively categorize customers, identify new clusters, and find new ways to think about "similar" accounts, locations, and customer groups.

It is particularly useful to identify groupings of before-and-after program participation accounts and locations. Similarities in building type, age, homeowner demographics, and the PAs' consumption data may allow the PAs to both prioritize the most attractive energy efficiency targets, and determine how similar customer types have benefited from participation in specific energy efficiency measures. It is possible that this type of

[6] Delivered fuels includes oil, propane, biomass, and others.

detailed targeting helps construct more meaningful narratives around items such as non-energy impacts (NEIs), which could make non-participants more interested in energy efficiency.[7]

This data will also be useful in any larger machine learning and data mining activities the PAs wish to undertake in the future, including the Bayesian Knowledge Networks discussed in the Customer Profile reports.

4.2.2 Behavioral changes from Energy Star benchmarking

The integration of the tax data with detailed building use codes, vintages, and Energy Star building scores allows the PAs to model a simplified Energy Star score for C&I accounts. The ability to model this information could help motivate high-scoring customers to register their build, could spur future savings by increasing awareness of energy consumption among participating customers, and could motivate customers that are close to Energy Star qualification but not quite over the threshold to engage their PA to install the necessary energy efficiency measures.

4.2.3 Location energy consumption, savings, and comparable sites

The wealth of building use information contained in tax, PA, and third-party integrated data allows for a very detailed segmentation of building types. Five years of billing and tracking data, linked across fuel, can be used to identify buildings that were comparable prior to energy efficiency installation, and to track changes that occurred after one of these buildings incorporated energy efficiency but the other did not. This information can help pinpoint specific sites as high-priority outreach targets for the PAs, and identify customer types that might represent a new targetable sub-population. It can also be combined with weather sensitivity data and energy use intensity (EUI) data to show which sites have high EUIs, what efficiency measures remain to be done, and particularly for residential customers, what their weather sensitivity is for measures like HVAC. Demographic data can also help micro-target geographic areas that are likely to have population groups more amenable to energy efficiency, without compromising customer information.

4.2.4 NEIs at larger industrial locations, locations subject to environmental regulation, or locations of environmental interest

A tool can be developed to use the EPA facility registry system and PA, InfoUSA, and tax data to identify sites that:

- Are high energy users relative to similar size and industry peers
- Have not installed energy efficiency measures in time-series data
- Have NEI regulatory or compliance considerations that impact their business

Such a tool could potentially target accounts or locations where a measure such as a new boiler could help a facility shift from a major pollution risk to a minor pollution risk, greatly decreasing compliance costs. Additional public compliance data could be used for targeting, as could state data such as boiler inspection data that includes boiler sizes and vintage.

The tool could also help pinpoint sites that could benefit from process measures or from spatial analytic tasks like proximity analysis on waste heat recovery. There is also the potential to identify community

[7] For example, a weatherization project may have the added benefit of reducing the exterior noise levels experienced in a home, making for a quieter and more comfortable building.

groups and areas of interest where there may be social or economic drivers beyond energy efficiency (e.g., high asthma rates, poor water quality, or availability constraints, etc.).

5 CONCLUSIONS AND RECOMMENDATIONS

The database enhancement project yielded a substantial amount of new data to support deeper insights into customers and locations while enabling more efficient project timelines and budgets. Expanding the tools, data, and logic models to the PAs' residential data greatly increased the impact, in a way that had not been envisioned at the start of the database enhancement project. PA feedback indicates that across residential and C&I sectors, this data has been highly useful in responding to strategy and implementation questions, evaluation needs, and regulatory information requests.

DNV GL has identified a series of recommendations for consideration —several of which the PAs have already implemented—to continue to grow the return on the investments made in the data enhancement project. These are as follows:

1. Annually rerun the tax integration process. This will ensure that the logic models continue to provide the PAs with the most recent building data. Accumulating time-series tax data in a database will also provide the PAs with a longitudinal sample of changes in building use, size, ownership, and value that may be useful in future analytical undertakings. DNV GL has already conducted one update cycle for the PAs to capture new data and integrate this with new customer accounts.

2. Continue to refine the logic models. Much of this project's effort went into the original build-out, testing, and revision of the logic models—all aimed at maximizing the balance between the highest possible match rate and the lowest possible number of false positives. Within the data it is likely that there are sub-populations with lower match rates or higher false positive rates than desirable. Typically, such sub-optimal rates are identified by looking at patterns in individual records within a known sub-population, assessing these patterns to see if they represent outliers or something systematic, and applying modified logic to the impacted records without adversely affecting non-impacted records. DNV GL works to improve the logic models though the annual data intake process to support these ongoing efforts and streamline the data integration process.

3. Integrate physiographic information with residential data. Successfully using the logic models with the residential data has revealed a unique opportunity, not present on the C&I side, for physiographic data integration. This type of information may be particularly helpful in understanding the subtleties of residential customer motivations. DNV GL has undertaken this type of data integration in the past in Massachusetts with very positive results. Physiographic data comes from data vendors, such as Experian and Axiom.

4. Continue to pursue ways to use this data. Ultimately, this project's value lies not in the data integration itself, but in all the additional analyses that it has made newly feasible. The data integrated for this project has already become part of many other planned and ongoing projects. The more it can be used, the more insights it can yield, and the more the PAs, evaluation contractors, and other parties can identify ways to improve it even further.

APPENDIX A: SELECT DATASET FIELD DEFINITIONS AND FIELD USE OPPORTUNITIES

This appendix presents a selection of fields from some of the integrated datasets, along with brief examples of how the fields might be used in analyses. This appendix is intended to provide stakeholders with an idea of the types and uses of the data integrated into the PAs' data ecosystem; it is not an exhaustive list, and not all the examples provided are planned as future analyses. The appendix includes details about the following data sources:

- American Community Survey Data
- Tax Assessor data
- InfoUSA data
- Dodge Players data
- EPA Facility Registry System data
- Zip Code Business Patters data
- LEED Buildings data
- Energy Star Buildings data

5.1 American Community Survey Data

DNV GL has not included a table for the American Community Survey (ACS) block group demographic data in this appendix due to the size and number of variables in that dataset. Interested stakeholders can find an archive of the ACS questions at https://www.census.gov/programs-surveys/acs/methodology/questionnaire-archive.html; the ACS homepage is located at https://www.census.gov/programs-surveys/acs/. The ACS data is geographic in nature and cannot be linked to any individual PA's account, nor can block group attributes be linked together (e.g., it is possible to identify that a block group is 50% non-English speaking, and that 10% of the block group commutes more than an hour to work; but we could not say that 10% of the non-English speaking population commutes more than an hour to work, or identify which PA accounts are included in the 10% of the block group that commute more than an hour). Nevertheless, the dataset provides a rich geotargeting tool that can help determine which of MA's over 5,000 block groups contain similar demographic profiles; it can also help isolate and understand areas of geographic variation in outcomes like savings, participation, and measure adoption.

5.2 Tax assessor data

Variable name	Short definition and use for select fields
Tax parcel key	A unique identifier to link back to tax parcels. This key field is important to ensure that as tax data is updated, the corresponding utility accounts can be updated to keep the data fresh.
Parcel zoning	May be a useful variable as high-level proxy for building type or geographic region type (e.g., industrial park, low density housing, etc.) that can support customer segmentation for models.
Last sales price	May be a useful proxy to identify distressed property where lower sales prices could indicate older equipment, building shell quality, etc., particularly when used in conjunction with assessed value and building vintage.
Building units	May help identify multifamily housing units and properties where there could be multiple owners or decision-makers.
Building style	Non-standard. Can be used as supporting information on NAICS, SIC, and other building type variables; can also be leveraged for accounts that do not have building type information associated with them. May be useful in model segmentation and classifying customers into similar groups. May also be useful in determining potential target measures or offerings.
Property use code	Standardized building use code. Can be used as supporting information on NAICS, SIC, and other building type variables; can also be leveraged for accounts that do not have building type information associated with them. May be useful in model segmentation and classifying customers into similar groups.
Assessed building value	The assessed value of the building on the parcel. May be useful for modeling. Lower assessed values may be useful (especially in conjunction with other fields) as a proxy for identifying lower income multifamily households based on deferred maintenance and surrounding property values; may also be useful as a normalizing variable.
Total record value	The assessed value of the parcel as a whole. May be useful for modeling, and (especially in conjunction with other fields) as a proxy for identifying lower-income multifamily households based on surrounding parcel values. May also be useful as a normalizing variable.
Building area	The building area. May be useful as normalizing variable for items like energy use intensity, and as a modeling variable.
Last sales date	The last time the parcel was sold. May help identify locations that are likely to have built up equity or that could be longer-term locations with willingness to take on longer payback measures. Alternately, may help identify recently purchased parcels where new ownership may be receptive to contact about efficiency investments, or used in conjunction with owner information to target locally active real estate investment groups.

Variable name	Short definition and use for select fields
Year assessed	The last year the parcel was assessed. May be useful in time-series analysis for a large population.
Owner name	The name of the person who owns the property. May be useful as a fuzzy matching variable for bringing together third-party datasets. It may also work as a proxy to identify instances where a parcel has many units but only one owner as potential commercial real estate, or to identify instances where the taxable owner of record does not match up with the person paying the utility bill as an indicator of possible split decision making between site occupant and site owner.
Building stories	The residential area in the building. May be useful as normalizing variable for items like energy use intensity, and as a modeling variable. May also help identify split-use buildings.

5.3 InfoUSA data

Variable name	Short definition and use for select fields
Affiliated locations	Verified records unique by location with the same subsidiary number as the selected record (branch records will not have a count). Helps determine customer-level linkages and set expectations about how many different addresses should be associated with the matching records.
Affiliated records	Verified records with the same subsidiary number as the selected record (branch records will not have a count). Helps determine customer-level linkages and sett expectations about how many total records in the data should be expected to align with parent IDs.
Company name	The name of the company that has been recorded as doing business at the location. Is an important variable for integration into the logic model to identify customers across PAs and to understand who and how data integrates for cross PA data linking.
Credit rating score numeric	Numeric version of the credit score. Potentially valuable data mining variable for firms, particularly small businesses, ability to invest in larger capital outlays on energy efficiency. May also provide useful input into understanding risk of securing lifetime savings from measures such as process due to business closure.
Employer identification number(s) (EIN)	Possible variable for grouping chains and franchises across fuels and PAs. Can be used in coordination with the other logic model variables to improve ability to match these customers while reducing the false positive rate.
Executive names	Possible variable for grouping chains and franchises across fuels and PAs. Can be used in coordination with the other logic model variables to improve the ability to match these customers while reducing the false positive rate.

Variable name	Short definition and use for select fields
Home business	Indicator flag for a business run out of a household. May be useful to better understand what types of measures small businesses are installing and to provide context on measures that may be metered and installed at a residential metered site (such as smart strips) but that are functioning as a C&I energy savings device. Potentially useful variable for targeting the micro/hard-to-reach C&I sectors.
IUSA number	Identification key for sites; can be used to track, update, aggregate, and share PA information on savings, etc. without relying on customer- or account-level information.
Location employee size actual	Detailed information on employment size within a business location. Likely a strong input variable (in conjunction with other firmographic information from PAs, tax data, and InfoUSA) for understanding individual location probabilities to engage in different energy efficiency measures; will help provide insight into the small business markets in conjunction with the other data integrated into the data ecosystem.
Location sales volume actual	Detailed information on sales within a business location. Likely a strong input variable (in conjunction with other firmographic information from PAs, tax data, and InfoUSA) for understanding individual location probabilities to engage in different energy efficiency measures. This is particularly true if there is a correlation using ACS demographic data and (when it exists) individual consumer data can be refined that might facilitate how customer characteristics and perceptions of a location's energy efficiency activities could help increase sales with greater promotion or visibility of the EE efforts (similar to how some retail locations post an interactive counter of how much solar power their solar system has generated and what this translates to for things like "trees planted." This will help provide insight into the small business markets in conjunction with the other data integrated into the data ecosystem.
Mailing address	Possible variable for grouping chains and franchises across fuels and PAs. Can be used in coordination with the other logic model variables to improve the ability to match these customers while reducing the false positive rate.
NAICS	Standard nation-wide NAICS code, can be complementary to existing PA data and other—in particular, tax use codes—fields for bivariate classifications. Data can be used in cluster analysis and categorization of accounts as well as understanding opportunities for measure savings by looking at what similar sized customers within the industry have done statewide. May also have value for the larger PAs to assess if certain measure types have gained traction in other service states that might be attractive customer offerings for a subset of customers in MA.
Own or lease	Indicates if the building is owned or leased by the customer of record. May be helpful in identifying CRE, and in conjunction with the tax data, help identify the ultimate decisions making entities for these organizations.

Variable name	Short definition and use for select fields
Parent IUSA number	Variable for grouping chains and franchises across fuels and PAs. Can be used in coordination with the other logic model variables to improve the ability to match these customers while reducing the false positive rate.
Rent expenses	Categorical variable indicating the rent range for the customer. Estimated. May be helpful in conjunction with the own/lease flag to refine CRE; and in conjunction with other variables on industry and PA tracking data, to better model and segment what types of measures are going into specific CRE management company locations. This may also be helpful in refining an approach to these CRE management companies by facilitating a detailed payback period prospective report that ties the PA energy efficiency expense to specific energy savings for the location, and a better representation of what the increase in rental income may look like.
Square footage	Complementary variable to the tax data; likely useful input into data mining and profiling models and as a proxy variable on scale and quantity of measures

5.4 Dodge Players data

Variable name	Short definition and use for select fields
Dodge project ID	This is unique ID assigned by Dodge to each distinct individual project record. May be useful for the cross-PA coordination of data.
Player Company	Project Player Company; can be leveraged to understand what types of new construction or retrofit activities and sectors individual companies specialize in. Possible input variable to probability models or network models for targeting specific measure offerings in coordination with contractors. May be useful in segmenting the new construction market actors.
Area	Dodge reports project square footage for newly constructed space. May be useful in conjunction with tax square footage to better understand system upgrade opportunities and what proportion of a structure has been renovated.
Primary work type	This is a final categorization of the project as either primarily new construction, addition, or renovation work to facilitate analysis with mutually exclusive categories.
Project start year	This indicates the approximate year that the project started building construction based on when the project was assigned a Dodge status of "Start." "Start" is assigned when the project contract is awarded, and work has begun or will begin within 60 days.

5.5 EPA Facility Registry System data

Field name	Short definition and use
Facility registry identifier	The identification number assigned by the EPA Facility Registry System to uniquely identify a facility site. This is the key for the online EPA query tool at https://www.epa.gov/enviro/frs-query-page. Allows for the aggregation of multiple accounts and locations within a larger EPA tracked site to be aggregated up to the EPA reporting grain and linked to the corresponding detailed EPA information tables.
Facility detail report	The URL address of the facility detail report for this facility. The report shows facility details, such as where the facility is located on a map, environmental interests and links to program systems, industrial activities, and affiliated organizations and contacts. Provides a link to third-party-maintained information for leveraging by evaluation, strategy, and implementation teams to ensure that PAs and stakeholders with access to the crosswalk can quickly reference the most recent version of the data without needing to rely on technical staff to pull from a data ecosystem. The HTML link includes information on enforcement activities, SIC, NAICS, critical contact names and alternate names of company for logic model matches, parent company names, and physical plus mailing address details. Additional links include detailed historical and future compliance actions for understanding emissions information including GHG and criterion pollutants that may have NEI benefits to local community and PA customer.
Last reported date	The most recent date the corresponding environmental interest data was reported to the Source of Data. In conjunction with facility and program status, this will allow PAs to capture active EPA sites and to understand sites that may have been historical opportunities, but for whatever reason have not recently been reported to EPA; a review of data suggests that when these sites have not closed, they may represent attractive opportunities even if they are no longer an EPA-monitored site.
Information system identifier	The identification number, such as the permit number, assigned by an information management system that represents a facility site, waste site, operable unit, or other feature tracked by that Environmental Information System. Provides the key within the different EPA monitoring systems to find information for the corresponding PA population of accounts. May be a particularly useful tool for data mining by identifying facilities with similar characteristics from the data ecosystem and assessing what types of measures, approaches, and engagements have translated to success for the PAs in the past.
Environmental interest type	The environmental permit or regulatory program that applies to the facility site (e.g., TRI Reporter, NPDES Major, NPDES Non-Major, Large Quantity Generator [LQG], Air Major, Air Minor). See interest type definitions at http://www.epa.gov/enviro/html/frs_demo/presentations/interest_types.pdf

Field name	Short definition and use
Active code	A code that indicates whether the environmental interest is active at the facility or site. Can be leveraged in data mining efforts and industry targeting for understanding similarities between sites and how this might translate to measures including inactive sites where data might otherwise be lost in time series tracking within the PA systems.
Alternative name	An alternative, historic or program-specific name for the facility site. Increases the likelihood that logic models and data mining will be able to integrate data across the disparate sources of information (e.g., linking tax to PA CIS to EPA); increases the universe of information that can be leveraged to profile customers. As an example, FRS ID 110000746845 corresponds to 14 different name variations across the EPA universe.
Affiliation type	The name that describes the capacity or function that an organization or individual serves for a facility site; associated with the contact name. May be useful outreach field for key programs - as example outreach to the air source major contract to better understand if new process or boilers that both save energy and reduce emissions are of interest. Can also support data mining for cross-customer identification when a parent company is flagged in the attribute data.
NAICS code	The 6-digit code that represents a subdivision of an industry that accommodates user needs in the United States. A facility may be performing more than one industrial activity and may have multiple values associated with it. Data can be used in cluster analysis and categorization of accounts and to understand opportunities for measure savings by looking at what similar sized customers within the industry have done statewide. May also have value for the larger PAs to assess if certain measure types have gained traction in other service states that might be attractive offerings for a subset of MA customers.
SIC code	The description of a subdivision of an industry that accommodates user needs in the United States. A facility may be performing more than one industrial activity and have multiple values associated with it. Data can be used in cluster analysis and categorization of accounts as well as understanding opportunities for measure savings by looking at what similar sized customers within the industry have done statewide.
DUNS number	The Data Universal Numbering System (DUNS) number assigned by Dunn and Bradstreet to identify unique business establishments. Can be leveraged to identify customers across locations and records in system and integrate back to the InfoUSA data as a complimentary information point.
EIN number	The unique tax identification number issued by the Internal Revenue Service to the employer. Can be leveraged in conjunction with other keys to aggregate and understand customers across accounts and PAs.

Field name	Short definition and use
State business identifier	The uniform business number assigned to an official business by a state. Can be leveraged in conjunction with other keys to aggregate and understand customers across accounts and PAs.
Supplemental environmental interest type	The supplemental environmental permit or regulatory program that applies to the facility site or the environmental interest at the facility site. For the purposes of FRS, supplemental program interests include state program interests and compliance and enforcement programs. May be useful for cluster analysis and segmentation as well as identifying recently impacted sites where NEI benefits may have greater near-term interest due to compliance activities that can complement the energy efficiency savings potential.

5.6 Zip code business patterns data

Field name	Short definition and use
EMP	The total number of employees in the zip code as of mid-March of the ZCBP year. Includes a noise variance. Useful for time series information on increasing or decreasing employment trends as a proxy for business climate and investment (detailed NAICS data with size range of employees can further target this beyond the base ZCBP data).
Empflag	Data suppression flag
ZIP	The zip code for the business patterns. Most granular level available to link into the PAs' data. Useful input variable for top-down models.
ap	The total annual payroll in the zip code as of mid-March of the ZCBP year. Includes a noise variance. Useful for time series trends on increasing or decreasing employment trends as a proxy for business climate and investment.
est	The total number of business establishments in the zip code as of mid-March of the ZCBP year. Includes a noise variance. Useful for time series data on increasing or decreasing employment trends to determine business climate and investment.
qp1	The total first quarter payroll in the zip code as of mid-March of the ZCBP year. Includes a noise variance. Useful for time series data on increasing or decreasing employment trends to determine business climate and investment. In conjunction with the annual payroll, can be used to refine within-year economic trends.
year	The analysis year for the ZCBP data

5.7 LEED building data

Field Name	Short definition and use
ID	The identification number assigned by LEED to uniquely identify a building. Links to the detailed project data indicating what points were secured in pursuit of the LEED certification (e.g., daylighting strategies, on demand ventilation, etc.)
Cert date	The certification date (if certified) for the project. May be useful variable for propensity modeling for measure adoption lag analyses.
Cert level	The certification date (if certified) for the project; may also indicate just that the project intended to pursue a certification (e.g., "registered") but never completed the paperwork. Possible useful variable for the geotargeting energy-aware locations or communities (e.g., school districts or military housing). May help PAs identify low-hanging opportunities to bump certifications up to the next level by installing energy efficiency measures—saving energy for the PA and garnering social recognition for the LEED building.
Is certified	Flag variable indicating registered versus certified buildings. May help PAs identify low-hanging opportunities to bump registered but uncertified buildings leveraging social recognition for the LEED building and generating energy savings for the PA.
Is confidential	Flag variable showing whether the record is confidential; confidential records have no accompanying data.
LEED system version display name	Indication of what LEED system the building went though
Owner organization	Indication of who owns the building. May be leveraged in the logic models to link up customers across buildings, and combined with the PA and tax data, better understand if certain customers have propensity to pursue these types of certifications at their facilities.
Owner types	Indication of the owner classification (municipal government, private, etc.). May help identify and geotarget stakeholder types with the propensity to pursue LEED in their jurisdiction. In conjunction with the tax and logic models, helps understand what types of measures have been installed at similar locations but not at the stakeholder sites.
Points achieved	Indication of the points achieved in the LEED certification process; should sum up to the total points in the detailed project table. May be used in conjunction with tax use code and square footage and PA industry code and energy consumption to build out model of similar but uncertified locations for targeting.
Project name	Name as recorded by LEED for the project. Usable in logic model to refine cross-fuel linking and to increase probability and quality of matches between different datasets. Can also be leveraged to identify groupings of customers pursuing LEED certification (e.g., "Stop and Shop #215") as a subset of larger population of customers from the tax data and PA CIS systems, with the potential to promote LEED or other labeling to other members of the group or to similar types of customers.

5.8 Energy Star building data

Field name	Short definition and use
B ID	The identification number assigned by the EPA Energy Star system to uniquely identify a building
Building name	Name as recorded by Energy Star of the building. Useable in logic model to refine cross-fuel linking and to increase probability and quality of matches between different datasets. Can also be leveraged to identify groupings of customers pursing energy star label (e.g., "Stop and Shop #211") as a subset of a larger population of customers from the tax data and PA CIS systems with the potential to promote energy star label to other members of the grouping or to similar types of customers.
Building owner	The entity or owner of the Energy Star-certified location. An important variable for integration into the logic model to identify customers across PAs and to understand who and how data integrates for cross PA data linking.
Building type	Standardized classification of building types per energy star. May be useful in developing benchmark proxies for similar building types (in conjunction with the tax and PA data) to understand if/how Energy Star labeling might impact existing buildings in the PA data via the EUI scores.
Floor space FT2	Indication of the floor space of the building when certified. May be leveraged in conjunction with tax use code and square footage and PA industry code and energy consumption to build out a model of similar but uncertified Energy Star candidate locations for targeting.
Label year(s)	Complementary variable with the rating score indicating when the building was last certified
Potential multifamily flag	Flag variable indicating if the Energy Star building is also multifamily
Rating(s)	The Energy Star score achieved by the participant building for the year certified. May be leveraged in conjunction with tax use code and square footage and PA industry code and energy consumption to build out a model of similar but uncertified Energy Star candidate locations for targeting.
Year constructed	The year that the building was constructed. Can be leveraged in conjunction with the building type and tax data to develop cluster model of similar buildings for an Energy Star targeted offering aimed at influencing commercial building operations and behavior.