USING EXPLORATORY FACTOR ANALYSIS TO GEOTARGET PERSONAS OF INTEREST

Rich Crowley, DNV, Portland ME; Riley Hastings, Eversource, Westwood MA; Steve Menges National Grid, Waltham MA; Autumn Snyder, Liberty Utilities, Fall River MA; and Robert Wirtshafter, Wirtshafter Associates, Inc., Chatham, MA

ABSTRACT

For both equity and efficiency reasons, rate payer-funded energy efficiency programs are becoming increasingly focused on attracting participants who have not yet taken advantage of program offerings. Evaluators are being asked to identify and describe customers and locations that have not participated in energy efficiency programs. Because participation can be defined as the building, account, or occupant there are ambiguities that make identifying these customers and locations challenging. Data commonly captured in customer information systems (CIS) can be limited and is often outdated. More detailed information can be expensive or impossible to acquire, legally restrictive to leverage, and difficult to maintain as an integrated field within CIS and energy efficiency tracking systems.

This paper shows how one state's Program Administrators (PAs) have overcome these data challenges using US Census Block Group demographic data, CIS data, other external data sources, and energy efficiency program data. An exploratory factor analysis (EFA) identifies Census block groups with shared socioeconomic factors and examines their relationship to participation rates. The result of this effort is the synthesis of 1000's of socioeconomic characteristics into descriptive factors that are used to generate five easy to understand natural groupings of the data. These five clusters are given a corresponding score showing their relationship to participation rates so the PAs and public stakeholders can better assess where customer populations of interest are located and if a particular geography is over or under participating relative to similar block groups.

The results of this analysis have been used by the PAs and public stakeholders for geotargeting and understanding spatial similarities and participation trends for communities of interest. The results are presented in an annual Residential Customer Profile report which provides stakeholders with the supporting context and details on the EFA. The customer profile report includes detailed maps to support outreach and engagement which allows the audience to visualize target communities.

In this paper we describe the data used for the EFA and how the EFA is conducted so that other utilities and stakeholders can conduct the same type of analysis to identify communities of interest in their own territories. We also include details on how the individual factors are developed, how different factors can be leveraged by stakeholders to assess and understand populations of interest, and how the top ten factors were refined into five block group personas that resonate with non-technical stakeholders. Finally, we present the graphics and maps developed though stakeholder working groups that the project team used to disseminate results from this analysis. We conclude the paper with a brief discussion of the lessons learned in the process of turning technical details into useful material for non-technical audiences in hopes that others will be able to build off our experience.

Abstractions of Monolithic Customer Groups Don't Have to Be the Norm

Energy efficiency program stakeholders – from utilities and regulator to implementors and community stakeholders – have long sought to understand which, and how, specific sub-populations of customers can be engaged via energy efficiency offerings. This is a simple question in concept, and one that in theory can be supported by the growing abundance and sources of detailed consumer data.

However, the often-overlooked challenge is individual customer economic and demographic characteristics data are rarely available at that individual customer level due to combinations of cost and third-party data licensing restrictions as well as data confidentiality considerations when using public sources of data. Customer level targeting can be further complicated when utility program managers may not have access to the same types of data as customer service departments and are less likely to be able to influence design changes to internal data stores and access policies in support of leveraging this data when it exits.

To maximize program reach and impact, program administrators (PAs) often adapt requirements and incentives to target customer sub-populations of interest. However, even with these program design incentives program administrators often still face challenges in getting these customers to participate. Customers are frequently considered in the abstract as monolithic group populations such as “renters” or “income eligible” when program designers or other stakeholders are limited in the information about specific customers, where these customers are geographically, what efforts have been made previously to recruit these accounts, and which customers have participated previously in this or other programs.

This paper presents an approach to use customer data on previous participation along with Census and other demographic and spatial data to better characterize the localized non-participating populations. It also addresses a common problem of balancing the mandate to execute cost effective programs with the inherent challenge that the more customized and targeted to individual consumers a program becomes, the higher the design and administration costs tend to be. Though the development of spatially bounded, descriptive customer personas[1] there is greater transparency into where, and why, certain geographies make more sense to prioritize. These insights also arm local stakeholders with the types of supporting information that they require to provide tailored support to key customer groups (Novie and Jarvis, 2016, R. Crowley, Senior Consultant, DNV pers. comm. Municipal Partnership Working Groups[2]).

Finally, the EFA method used in this paper also provides a pathway to reducing another persistent obstacle on the identification, description, and ultimately recruitment of non-participant customers: utilities’ obligation to protect the confidentiality of individual customer data. The approach we used can meet utility and stakeholder desire to expand service offerings that reach the largest swath of customer, using freely available tools and open-source data, while balancing stakeholder desire for granular insights into localized populations[3] with utility regulatory and legal obligations.

Using Census and PA Data to Characterize Block Group Households

The Massachusetts PAs have tackled developing descriptive customer persona groups tied to geographically discreet areas through the fusion of their program and customer data with open-source, economic and demographic data from the US Census Bureau’s American Community Survey (ACS).[4] ACS

---

[1] Personas are fictional characters used to represent a user type. In our case persona help explain a participation pattern.

[2] This theme has been reaffirmed though our engagement in multiple data support working groups after the completion of the profile report. Most recently this was articulated in a series of municipal stakeholder support groups around a soon-to-be-released series of interactive municipal geotargeting maps leveraging Google earth.

[3] These needs can be extremely varied. In MA examples have included but are not limited to sustainability department building GHG inventory road maps, community support networks seeking to better identify and engage their target populations, and implementors seeing to prioritize outreach in a way that most efficiently focuses limited resources.

[4] The MA PA’s also use additional, state specific data including economic development, tax data, and HUD data; these add additional value but, as they are state specific, we did not include them in the inputs for this analysis, and they are not elaborated on for this report.

data can be accessed in excel, SAS, or via geographic datasets and is comprised of a mix of household and population counts for defined groups of data in geographic areas as small as the Census Block-group[5] as well as continuous variables like median income or average commute times. Data is available at different granularities, with a larger number of attributes available as the spatial granularity is reduced to larger areas to protect customer from reverse engineering of data. At the most granular level – the block[6] – generally only household counts and population data is publicly available. The block group level is the most spatially detailed that also includes a large amount of demographic data. The tradeoff though is that this data is a 5-year rolling window of responses rather than a snapshot of a single point in time which can complicate analysis for time series phenomena (gentrification as an example).

On the utility side, measure installation data from tracking systems and rebates provides the input for total savings and a way to calculate participation. Participation can be tracked at the physical address, building (in the case of multifamily or campus locations), or at the account level. Challenges with each of these data grains have been previously discussed in greater detail in the annual MA Customer Profile reports, as well as in both the MA Residential, and Commercial, Non-Participant reports (DNV 2021, DNV 2020, DNV 2020a, DNV 2019). Account numbers are usually a required key to link to the rebate data and billing systems; however, account numbers can change over time as customers move or start and stop service. Building premise can be an appropriate measure of participation because most measures installed by the utility programs tend to remain in place. For this reason, we also use account number as means to connect to a standardized physical location in the utility billing system.

This linkage allows us to better reflect when an account that has not participated is at a location that already has a new boiler, water heater, and lighting from a prior customer account. In these instances, the current account is a non-participant but the location they are at has already participated during the measure lifetime window. The implication is that not all non-participants may represent an equal future opportunity, even when the current customer of record is under the impression that they personally have not been engaged with an energy efficiency program. The billing data from the utilities is also used when we want to look at savings rates, or consider alternate weightings for participation (e.g., one large master-metered multifamily account may make up a large fraction of the total consumption in the block group).

The PA customer billing and tracking data is aggregated up to the block group level as it both presents a localized grouping of customers while also allowing the ACS data to be leveraged without the need for additional processing steps. The result is the ability to make statements like, "Block group 12345 contained 10 customers consuming a total of 500 units of energy. Three participated in a PA program and saved 50 units of energy. Thirty percent of block group X participated, with savings amounting to 10% of the block group's total energy consumption." Once the PAs' billing and tracking data have been aggregated to the block group level, it becomes possible to make observations such as "Block group 12345 is similar at the population level to block groups A, B, and C with X% income eligible households, Y% renters, and Z% limited English proficiency.[7]

The value of open-source data, nationally maintained files like the ACS data, should not be underestimated. Although there is certainly a transactional cost to using the data, it provides utilities and

---

[5] Block groups are the most granular geography that is publicly available for the ACS data. They are generally defined by the US Census Bureau to contain between 600 and 3,000 people. They are non-overlapping geographic areas allowing for characterization of customers in the geographic space they exist in.

[6] Readers are likely to be most familiar with census blocks as the base input for the decadal reappropriation of US House districts that is currently ongoing in many states.

[7] It is important to note that in absence of additional statistical processing we cannot apply block group ratios to sub-populations of PA customers without committing an ecological fallacy—that is, inferring something about specific individuals based on inferences from the populations to which those individuals belong. While it is accurate to say that the relationship between income and participation is positive at the block group level, it would be inaccurate to state that within the block group, higher-income homes are participating at a higher rate.

stakeholders with a consistent, time series dataset that has already been captured and vetted for statistically representative accuracy. Additionally, since the data is provided at non-overlapping spatial grains (block groups, tracts) it is a relatively low lift to combine data for different territories and analysis grains. This provides flexibility to look at things including service networks, franchise territories, and community groups of interest individually. The data in this format can also be used as an input for more advanced analysis and comparison grouping.

The block group data can be presented "as is" in scatterplots, which provide a useful tool for quickly looking at generalize relationships between things like a block group's participation rate and the amount of income eligible households in the block group. Integration of this data with geographical tools such as Google Earth further allows users to react to visual clues and prompts in the data and facilities a means for institutional knowledge that might not be readily reflected in databases to be volunteered and brought to bear in support of increases customer engagement.[8]

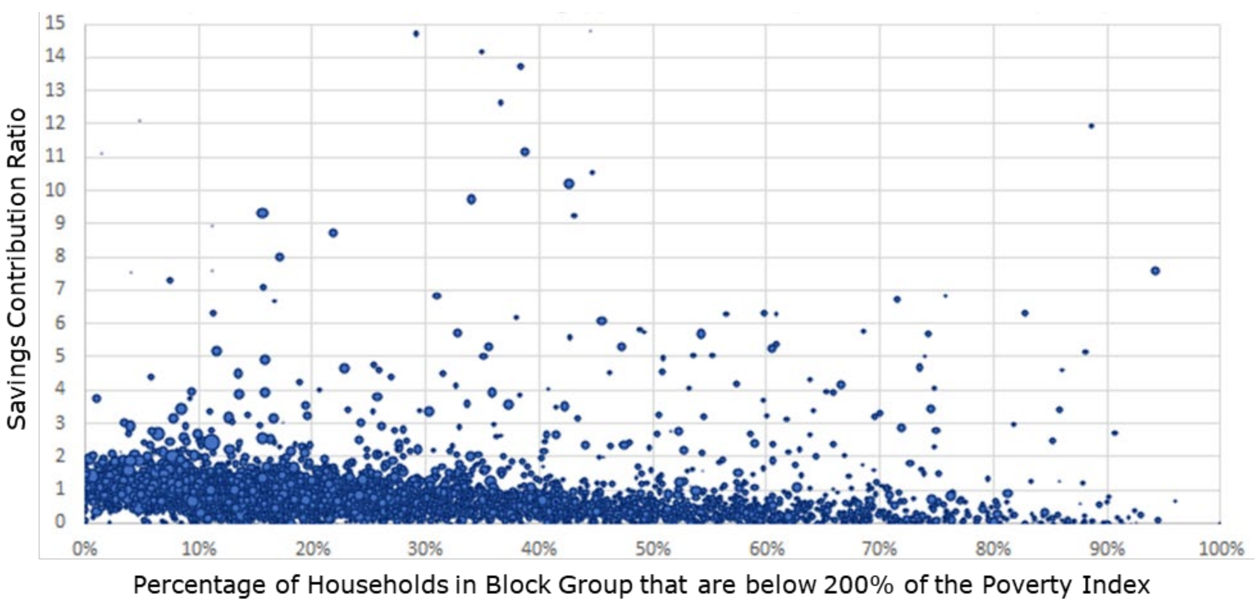


Figure 1. Block group ratio of households with annual incomes below 200% of the poverty index by 2017 electric savings contribution ratio[9]

Natural clusters and patterns in the data can quickly be identified even without computational steps. This type of information can be leveraged by stakeholders to inform decisions and engagement strategies without the need to conduct additional analysis – allowing stakeholders to take advantage of their knowledge of local communities without ever needing individual customer data. However, there are

[8] A recent example of this was a municipal stakeholder that had a family friend had a storefront location in a block group matching one of the customer profile's persona groups that had low participation and a high share of priority household characteristic both within, and neighbor to, the block group. An implementation vendor at the meeting was looking for an avenue to conduct some table outreach in high traffic areas and the real time connection of need versus institutional resources resulted in a verbal handshake and plan to get resources on the ground at the storefront via a trusted contact's introduction. Other MA studies have found this to be a key way to increase non-participation for the personas of interest (Illume, 2020).

[9] The contribution ratio is a unitless number reflecting the block groups share of overall contribution to statewide savings totals relative to the block groups contribution to overall statewide consumption. 1.0 indicates the share of savings, and share of consumption, are perfectly in proportion; caution should be used extending this to assume 1.0 indicates a subjective measure of "good" or "poor".

limitations in stopping the analysis at this stage. Chief among them is that many of the household variables are correlated with each other.[10] This is observed in the scatterplot; some block groups with higher poverty index ratios seem to also have higher savings contribution ratios leading to the question "are these block groups truly apples to apples comparisons?"

If we could capture these relationships, we could generate a "new" factor that reflected these relationships in a single descriptor. This would allow us to integrate more information into the models – though at the cost of more abstraction and less transparent math – and better segment sub populations on multiple dimensions.[11] To accomplish this the MA PAs used an exploratory factor analysis (EFA) of ACS and PA data at the block group level.

Conducting an EFA to Further Understanding of Household Participation

Exploratory Factor Analysis is used to identify the common factors – which are comprised of measured attributes, in this case ACS block group socioeconomic variables – that explain the order and structure among those measured variables. The factors are taken to represent the unobservable variable that reflects the correlations between the different observed variables from the input set (Watkins, 2018). EFA allows for each variable to be reflected in multiple factors and seeks to identify where those relationships are strongest in the data. This multi-factor relationship is a major reason for using the EFA approach in this analysis, rather than something like a Confirmatory Factor Analysis or other models where variables are limited to a single factor. The ACS data provides 1,000's of variables to work with and we know that there are interrelationships between them.

Starting the analysis with a limited, pre-determined set of variables is not necessarily a bad thing – particularly if you want to know about the relationship those specific variables have with the target function. However, there is value to looking at the larger set of data when you are trying to answer "what things seem meaningful" as a starting point to understanding why and how to leverage them. As a thought exercise example: knowing that commute distance has some explanatory relationship power to participation means we should actively encourage discussion about why and what could be a driver of that explanatory power. Perhaps it is that participants who commute primarily via mass transit are not as likely to get to a store where upstream programs are sold, let alone cart home a new water heater. Recognizing that commute distance has some explanatory power will improve our understanding of participation dynamics even though commute distance was not a focus variable of our initial question or prior program design considerations.

At the core, the EFA recombined several hundred ACS variables into multiple different factors that captured things like family structure, age dynamics, and economic characteristics for the block groups based on the PA and ACS data. However, reaching this point does require thoughtful consideration as well as preprocessing of the data.

Developing the Factors

The set up and execution of the EFA was broken into four broad steps:

1. Numerical counts in the ACS data were transformed into proportions by dividing the count of each class of data in a series (e.g., total number of households with income under 10k) against the parent ACS variable (e.g., total number of households for which income was

[10] One example is that income and age; educational attainment and income is another.

[11] E.g., all renters in block groups are the same, but real-world evidence suggests that rental homes in a working area of a city are not the same as vacation rentals, which are not the same as student housing. In these cases, income, age, and educational attainment would all help us to better classify the block groups.

determined). Non-count data (e.g., block group median income) were normalized between the values 0 and 1 using a range standardization.

2. After step one, the 3,588 census variables that remained as inputs were run though a correlation analysis against participation as a pre-processing step to reduce the number of inputs and focus only on those that had at least passing correlation with participation. A cut off value of correlation > 0.20 was set which resulted in 500 variables being kept.

3. To further narrow the list from 500 variables and simplify processing, DNV dropped variables for which data was missing for large swaths of block groups. This reduced the number of input variables from 500 to 339 variables correlated with participation.

4. Finally, these 339 variables were run though the factor analysis to capture the interrelation between the variables. The EFA was run with oblique varimax rotation with the a priori assumption that the factors would be correlated with one another, and our analysis wanted to obtain these interrelated elements as the factors rather than remove them from the analysis. A minimum eigenvalue[12] – the amount of variance that could be explained by the factor – was set to 1.

The outcome of the EFA with the eigenvalue floor was a set of 32 factors. A scree plot was used to visualize the factors and assess if there were natural break points to further restrict the number of factors used in developing the customer persona. The number of factors that are kept is an important decision point in the analysis: more factors mean greater variance explained but at the cost of greater complexity. Natural break points were identified at factor 7 and again at factor 11. In seeking to balance simplicity while also reflecting the factors that aligned with stakeholder priorities, and in consultation with the stakeholder teams, the breakpoint was set at factor 10. This was due to the combination of proximity to the natural break at 11, and because the key elements of factor 9 (largely educational attainment) were flagged as an area of interest for some stakeholder groups.

DNV ran a correlation of the factors produced from EFA and interpreted the factors through the significance of individual Census variables to each factor. The result was three broad categories of the 10 factors with a dominant thematic attribute as identified for each factor in Table 1. This provides stakeholders who are looking for the simplified, top line results with a way to identify the factors and the corresponding maps and personas that matter most to them.

Table 1. ACS factors and their dominant attribute from the EFA

Category	ACS Factor	Dominant Attribute
Family Dynamics	Factor 1	Single and Non-family[13] households
	Factor 10	Married Families
Age Dynamics	Factor 3	Retirees
	Factor 4	Younger, Group Living
	Factor 8	Old and Middle-aged Workers
Economic Factors	Factor 2	Renters

[12] An eigenvalue is a measure of the amount of variance an individual factor accounts for independent of all the other factors (Watkins, 2018). This variance can be plotted onto a scree plot to assess where there are substantive changes in the variance explained which in turn can help set a transparent and replicable decision point on the number of factors to be included in an analysis.

[13] A non-family consists of a householder living alone or who are living together but not related, such as a roommate or non-married couple,

	Factor 5	Income Eligible
	Factor 6	Minority Groups
	Factor 7	High Income Earners
	Factor 9	Education

Within each fact there are many different ACS attributes in addition to the dominant attribute listed. Table 2 presents a more detailed look at some of the types of factors and the relationships that ended up in factor 5 (income eligible) within the larger economic factor group. This presentation grain strikes a balance for stakeholders who may be looking for more context, but who do not necessarily need all the supporting math and details. A policy focused stakeholder looking to understand how regulations impact multiple criteria is the target audience for the additional text detail, while a business stakeholder looking to use what already exists as quickly as possible would be the target audience for the top line results of the prior table.

Table 2. Contributing ACS variables: EFA factor 5 example

Factor 5	Income Eligible	People 20-64 years that are non-veteran, not disabled, and not in labor force with income below poverty level (+); Workers who walked to work (+); Renter occupied units of 5 or more in structure (+); Renter occupied units with no vehicle available (+); With Medicaid/public coverage (+)

All Massachusetts census block groups have their own corresponding values (positive or negative) for each of the ten factors shown in above. This information can be used by stakeholders to focus their outreach efforts on specific sub-populations and geographies without having to go to the level of a full descriptive persona for the block group that leverages all the individual factor elements.

As an example, if a stakeholder wanted to focus on income eligible customers, factor 5 would be a logical place to start as many individual variables within this factor relate to income. Within factor five, the stronger the factor loading for an individual block group is the higher relative interest this block group would be for the stakeholder. By integrating the data into a geographic dataset and map, the stakeholders can see where these customers are occurring, if there are areas of statistically higher clusters in spatial proximity, and even (by integrating Google Earth files) the types of social support networks that may be in proximity to these areas. An example of this, albeit with a simplified set of data underlying the factor analysis, occurred at a stakeholder meeting on the MA Municipal Mapping project. A preponderance of Spanish language support resources, including a community heritage center, helped connect the dots for the community partner and implementor that deploying additional Spanish speaking staff in this area could be an effective way to connect with the community. Additional scatterplots and maps, including where clusters are located, can be found in the 2013-2017 MA Residential Customer Profile Analysis cited in the resources section.

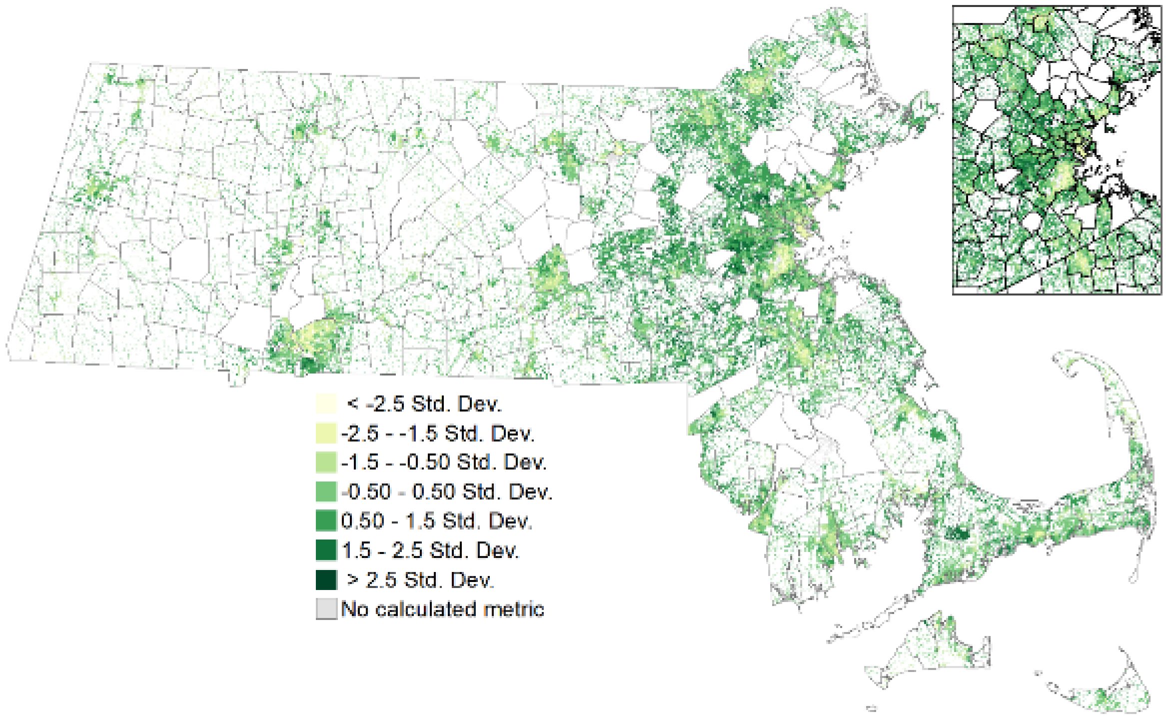


Figure 1. Income eligible (Factor 5) standard deviation from statewide mean factor scores by block group

Refining the Factors and Developing a Block Group Persona

The scatterplots, clusters, and individual factors from the EFA present a wealth of decision support information across a series of increasingly granular results to support the needs of different stakeholder audiences. However, the results can still feel technical and dry for audiences. Packing them into a descriptive persona can provide that final delivery that – in a word or two – sticks with larger audiences and is intuitively relatable.

Table 3 shows a detailed summary of the translation of ACS factors (F1 through F10) into five descriptive personas, along with summary stats on max, min, and mean participation rates. The importance of the final choice of words for the cluster persona should not be overlooked and was one of the more passionate areas of discussion in the working groups.

This table is the synthesis of the detailed numbers that a data analyst might look for, along with the connection to the policy ad business analyst's individual factors and elements; but packaged into that relatable persona that will intuitively resonate with a broad audience. The higher the factor score, the more that the factor contributed towards the type of customers found in the block group, and the greater weight was placed in ensuring that the descriptive persona reflected this type of customer. As an example, factor 4 (younger, group living) had a very high mean factor score at 1.55 within the persona for "students and city professionals" and so the ultimate descriptor for this persona placed a higher value on reflecting that subset of customers.

Actionable intelligence with geographic block group personas

The outcome of the EFA and supporting analyses and working groups was a plethora of data, as well as ideas of additional value that could be derived from this information.[14] Figure 2 shows the distribution of the five persona across the state. Figure 3 presents an example of one of the five "one page persona" sheets that was developed to give users a quick understanding of each persona.

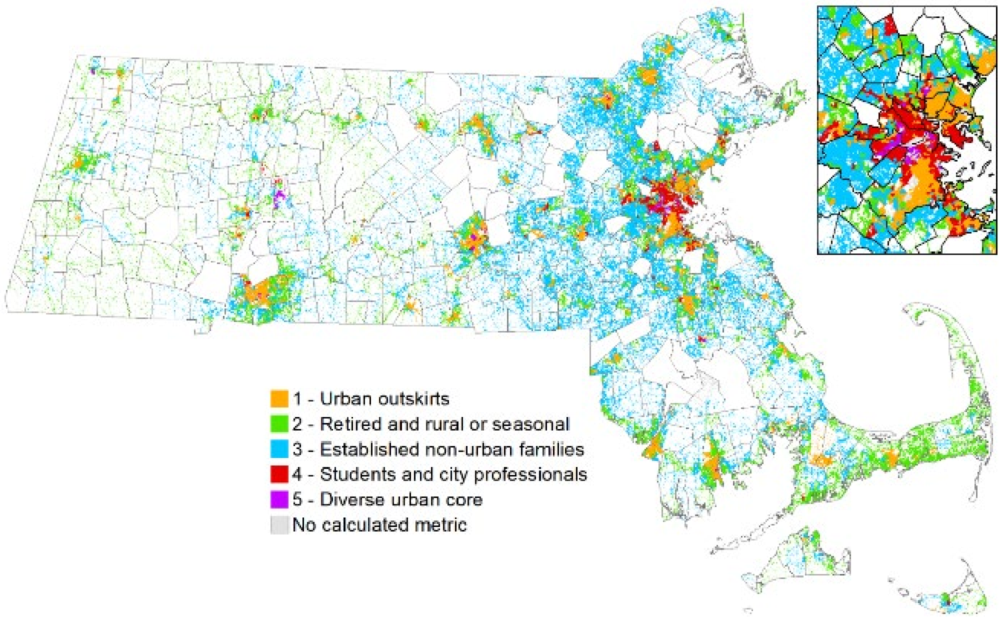


Figure 2. Statewide block group cluster personas and geographic dispersion

The utility of this type of graphic was readily apparent in conversations with stakeholders. Some of the value could be found in providing greater quantitative context to things that utilities and stakeholders know from experience – for example that populations on Cape Cod generally skewed older and that seasonal housing was a factor. Reaching older, and seasonal, customers may require different approaches than perhaps are appropriate for areas with more students or families. A networked enabled smart thermostat might be intuitive and attractive to a young person and unappealing to someone older. The message that resonates there might be more along the lines that that once the thermostats is programmed… it can do all those web enabled internet things when you are not at the home… but you can also just ignore it and it will work to.[15]

---

[14] Some of these items have already been integrated into other MA work including a non-participant customer profile effort, interactive block group level mapping tools for municipalities and an online data dashboard with a socioeconomic analysis lens. These items can be found at https://www.masssavedata.com/public/home

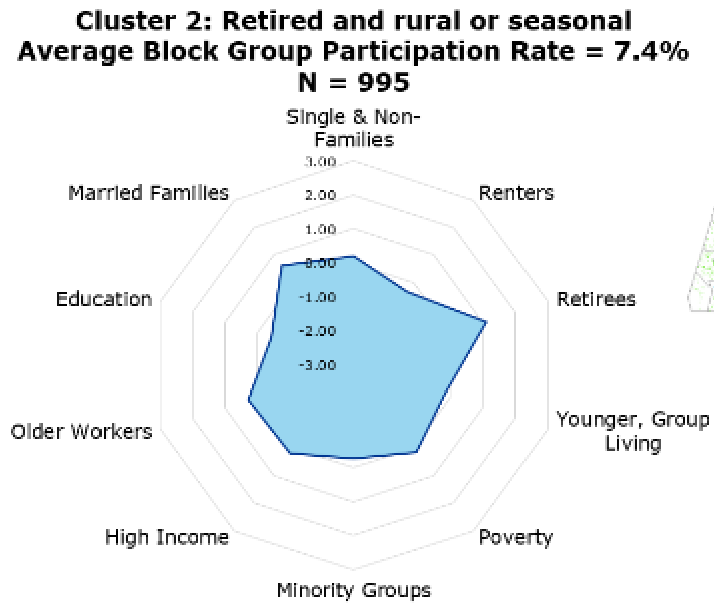
[15] Here at least one member of this paper team is speaking from experience trying to help their parents understand what an internet enable smart thermostat is, and ultimately only selling them on it when they were assured that the "up and down" arrows functioned just like the old mercury float that needed replacing.

Table 3. Translation of ACS factors (F1-F10) into clusters, detailed

Cluster	Statistics	Participation Rate	F1: Single/Non-family HHs	F2: Renters	F3: Retirees	F4: Younger, Group Living	F5: Poverty	F6: Minority Groups	F7: High Income	F8: Older Workers	F9: Education	F10: Married Families
1. Urban outskirts	Maximum	80.4%	4.63	3.57	1.69	3.33	2.59	5.85	2.83	1.97	2.87	1.72
	Mean	4.5%	-0.10	0.99	-0.49	-0.31	-0.97	0.63	-0.28	-0.75	-0.16	-0.30
	Minimum	0.1%	-3.01	-2.16	-2.33	-2.58	-3.67	-1.39	-3.37	-2.02	-2.78	-3.70
	Observations: 1,206											
2. Retired and rural or seasonal	Maximum	34.0%	3.42	2.48	5.85	1.91	2.32	2.31	3.45	2.47	3.94	1.82
	Mean	7.4%	0.20	-0.36	1.11	-0.19	0.14	-0.27	0.20	0.25	-0.44	0.61
	Minimum	0.9%	-1.81	-3.10	-1.39	-2.05	-2.89	-1.46	-3.16	-1.52	-2.83	-1.72
	Observations: 995											
3. Established non-urban families	Maximum	50%	2.51	2.54	2.22	2.50	2.76	1.87	3.61	2.98	4.84	1.68
	Mean	8.7%	-0.43	-0.65	-0.01	-0.26	0.65	-0.43	0.55	0.65	0.78	0.37
	Minimum	0.3%	-2.29	-3.07	-2.12	-1.51	-1.95	-1.75	-2.22	-1.62	-2.00	-1.78
	Observations: 1,515											
4. Students and city professionals	Maximum	29%	3.18	2.53	1.91	6.61	2.24	1.57	3.41	1.84	1.56	1.18
	Mean	3.4%	0.84	0.34	-0.80	1.55	0.05	-0.40	-0.96	-0.58	-0.88	-1.30
	Minimum	0.1%	-2.01	-2.22	-2.56	-1.23	-2.70	-2.50	-3.73	-1.74	-2.75	-3.95
	Observations: 484											
5. Diverse urban core	Maximum	38.5%	5.36	1.64	3.41	6.20	1.90	7.78	0.53	1.23	2.87	1.31
	Mean	2.4%	1.47	-0.18	-1.06	1.92	0.21	3.18	-2.15	-0.39	-1.16	-1.57
	Minimum	0.1%	-0.97	-4.08	-2.41	-1.57	-2.10	0.39	-4.56	-1.47	-2.71	-3.37
	Observations: 112											

Retired and rural or seasonal block group persona

The Retired and rural or seasonal cluster is mid-sized (N=995 block groups) with the second highest participation averages relative to the other four clusters. Retirees living in rural areas of the state or on the Cape characterize these block groups, but other factors are still present though at more suppressed levels of impact overall. Geographical variation in individual factors such as income by Otis Air Force Base and the Barnstable Airport can help further identify where micro-areas of customers exist from home elements like income are a greater factor and who may therefore qualify for measure under income eligible or renter programs.



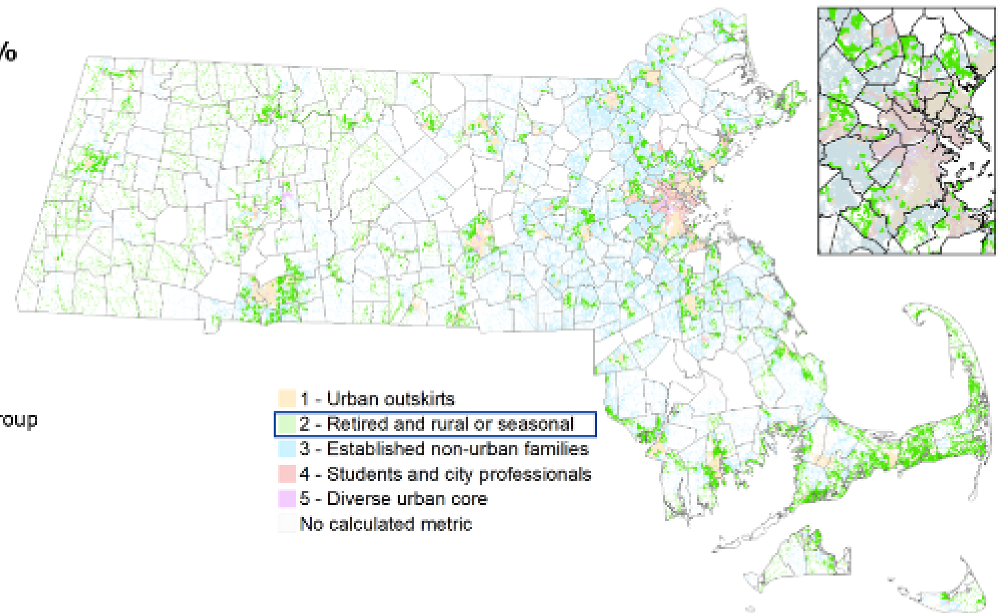


Figure 3: Descriptive Page for Cluster 2

The ability and value of being able to better segment customers and geographic areas within towns was also readily apparent. Using the Boston are zoom in from Figure 2 as an example, stakeholders could observe that that although rental housing and income are often a challenge in the city there are very different sub-regions of customers. Students and city professionals could be found in the core city area and along the waterfront while housing to the south of Boston still consisted of higher renter populations but with lower incomes, less younger non-family group living, and higher population diversity where multiple languages might be encountered. This type of insight both complemented existing stakeholder knowledge of where income eligible customers were likely to be found, while providing an important differentiation in the types of renters and housing dynamics that would be encountered in the otherwise ubiquitous rental markets of the Boston area.

Although not explored in depth in this paper, there is also great value to stakeholders in being able to identify comparison, or like, block groups in other areas of the state. This is particularly true when coupled with tax parcel data on home vintages, types, and size. The identification of these comparison populations provides stakeholders with a starting point to leverage statewide networks and engage with providers outside their core service area to understand what measures, approaches, and techniques have been successful in these comparable areas.

On the utility side, access to time series data on participation, consumption, and measures installed can provide a powerful data set for more tailored marketing, such as suggestions of measures that neighbors have successfully installed. This provides an avenue for concierge type customer outreach, but at scale and without having to purchase that very expensive and legally restrictive data.

Considerations and Lessons Learned from the EFA Process

An EFA can provide a lot of useful and informative data for a broad population of stakeholder, but the process of accomplishing this is not without its hurdles. Chief among them is finding the balance between an intuitive and actionable presentation of the data while securing trust and buy-in for a mathematical tool that is not always the most intuitive or transparent to the average stakeholder. This was largely overcome in MA though the engagement of stakeholders in working groups, and some very detailed feedback and reworking of the presentation of material into a comprehensive customer profile document. The sooner in the process that stakeholders can be brought into the conversation and presented with the value propositions that the analysis will provide – rather than the mechanics of how the analysis' math works – the greater the likelihood for that trust, and subsequently the greater the enthusiasm towards using the results.

Next, there must be give and take – especially on the pieces of the analysis that, while data informed, are more subjective. As mentioned earlier, the naming for the factors, and the block group personas generated discussion nearly as intensely as did the mathematical choices made by the modeling team. As we have seen time and again as a society in the last two year there is huge power in words: collaborative and thoughtful compromise may be tough but the value in not having labels and classifications that groups may find off-putting to point of dismissing the data is a critical step.

On the technical process side, it's imperative not to confuse free and open-source data and tools to mean that there is not a cost to leverage the information. As laid out in the factor development, a large amount of data processing and reduction was done to arrive at the outcomes, and this is not something that could be easily done without at least some specialized tools.[16]

[16] The team conducting this work leveraged SAS and PowerBI for the factor analysis, the clustering, and the data management with ESRI software used to generate the geographic outputs. Open source options including QGIS and Orange data mining can also be used to conduct an EFA.

Finally, the EFA project was initially undertaken intending to support the MA Residential Customer Profile project; at the time the graphics were static images, charts, and tables in a paper report. Static images are a starting point with the EFA but should not be the end game – if stakeholders want to take advantage of the detailed block group outcomes. Over the course of 2020 and 2021 the MA PAs have continued to build out interactive tools for stakeholders to access the types of data underlying the EFA to the point that the work presented here could all be updated by any party so inclined to pull down the information from the MA interactive dashboards and maps. The material presented here continues to be integrated into tools like interactive maps embedded in Google Earth and the feedback from stakeholders leveraging these tools has been overwhelmingly positive. Utilities, or stakeholders, undertaking profile work like this EFA should absolutely consider early on how to best disseminate interactive data in a way that is accessible to target audiences.

References

DNV, 2020. "Residential Nonparticipant Customer Profile Study." https://ma-eeac.org/wp-content/uploads/MA19X06-B-RESNONPART_Report_FINAL_v20200228.pdf

DNV, 2021. "2018 C&I Customer Profile Study Report." https://ma-eeac.org/wp-content/uploads/2018-CCPS-MA20X01-B-2018-Data-Insights-Final-Report-2021.02.03.pdf

DNV, 2020a. "Commercial and Industrial Small Business Nonparticipant Customer Profile Study." https://ma-eeac.org/wp-content/uploads/Final-MA19X11_B_SBNONPART-Report-20200415-1.pdf

DNV, 2019. "Residential Customer Profile 2013-2017." https://ma-eeac.org/wp-content/uploads/FINAL-2013-2017-RCPS-core-report_20190701.pdf

Illume, 2020 "Residential Nonparticipant Market Characterization and Barriers Study." https://ma-eeac.org/wp-content/uploads/MA19R04-A-NP-Nonpart-MarketBarriersStudy_Final.pdf

Navigant, 2018. "Census of Massachusetts Multifamily Buildings." https://ma-eeac.org/wp-content/uploads/RES43-Final-Report-2019-05-31.pdf

Novie, Alex and Jarvis, Daniel, 2016. "Sociodemographics and Small Businesses: Benchmarking Historical Participation Patterns across SBDI Programs." https://www.aceee.org/files/proceedings/2016/data/papers/4_661.pdf