Virtual Home-Auditing: A Statistical Investigation Using Publically Available Data on Gainesville, FL, Building Stock

Submitted in partial fulfillment of the requirements for

the degree of

Doctor of Philosophy

in

Engineering and Public Policy

Enes Hoşgör

B.S., Geological Engineering, Middle East Technical University
M.A., Energy and Earth Resources, The University of Texas in Austin
M.S., Engineering and Public Policy, Carnegie Mellon University

Carnegie Mellon University
Pittsburgh, PA

December, 2013

Abstract

Energy efficiency (EE) and energy conservation today are recognized as the low-hanging fruit of energy sources. However, the potential benefits of energy efficiency are often unrealized due to market failures and market barriers. The overarching objective behind my work is to merge publicly available data, e.g., property tax dataset for physical properties of households and voter registration data set for demographic household properties, to build statistically significant insight on energy efficiency and consumption for a group of households (n=7,091) in Gainesville, FL. This will explore and try to verify the concept of an energy efficiency reservoir. Absence of data is one of the biggest barriers to information flow and efficiency deployment that I aim to overcome in my thesis. The generated insight will be provided to different efficiency stakeholders, e.g., electric utilities, homeowners, contractors, home energy performance product providers, for them to implement their investment strategies in an informed manner.

This dissertation is dedicated to my parents, Ayşe Kadriye and İzzet Hoşgör, for their eternal love and support.

Acknowledgments

Funding for my first three years of research was provided in part by the Carnegie Mellon Electricity Industry Center and the Center for Climate and Energy Decision Making Center, through a cooperative agreement between the National Science Foundation and Carnegie Mellon University (SES-0949710).

I would like to thank Nick Taylor, Hal Knowles and Pierce Jones with Program for Resource Efficient Communities (PREC) at University of Florida for sharing the utility and publically available data that underlie my work.

I am thankful for the supervision provided by Arthur A. Boni, H. Scott Matthews and William F. Eddy, who make my doctoral committee.

Finally, I wish to extend my deep gratitude to my advisor Paul S. Fischbeck, for his kind mentorship, support and patience.

Table of Contents

List of Figures

List of Tables

1. Introduction

Today, energy efficiency (EE) and energy conservation are recognized as the low-hanging fruit of energy-reduction strategies. However, the potential benefits of energy efficiency are often unrealized because of market failures and market barriers. The overarching objective behind the work in this dissertation is to merge publicly available data (e.g., property tax datasets for physical properties of houses and voter registration datasets for demographic house properties) with utility-consumption histories and then to construct statistically significant predictive models for energy efficiency and energy consumption for a group of houses (n=7,091) in Gainesville, FL. This paper quantifies the concept of an energy-efficiency reservoir, the potential benefit of implementing EE programs.

To date, EE predictive models have either been handcrafted for individual structures or based on broad sets of average characteristics (e.g., aggregate data from the Department of Energy's Residential Energy Consumption Survey (RECS)). In contrast, the generated models from this dissertation provide different stakeholders (e.g., electric and gas utilities, homeowners, contractors, home-energy performance product providers) with valuable tools for determining where investments should be made to improve the efficiency of the housing stock for a broad regional area.

Three studies constitute the core of this dissertation all of which use the same aforementioned data set (n=7,091) that includes 1) monthly utility usage data for electricity, natural gas and water for 36 months from 2009 through 2011, 2) property tax records that have baseline information on the physical characteristics of the house structure (e.g., square footage, year built, building value, and number of bedrooms of the houses), and 3) voter registration records for the demographic profile of houses (e.g., lower bound on the number of occupants, number of adults, number of Republicans). The three studies are written as independent stand-alone working papers.

The first study (Chapter 2) investigates the statistical descriptive power of publically available information (i.e., property tax records and voter registration records) for modeling utility usage. In this paper, we show the importance of distributing monthly utility data to daily reads spread across the billing period. Simply assigning the monthly bill to the month of the bill distorts the consumption record especially for months where temperatures vary from the first to last day of the month (e.g., months in the spring and fall). Using these adjusted consumption values, we explore two different regression modeling approaches: 1) separate regression models for each month and 2) a single model with monthly dummy variables and monthly temperatures. The second model provides much better predictive power. The motivation behind this study is to build a framework that can be used by utilities to plan for monthly changes in demand with respect to demographic and structural characteristics in their service territory. The predicted monthly utility usage values can also be translated into a energy-use intensity per squarefoot which can be used as a first-cut metric for energy-efficiency planning and targeting for utilities who are mandated by state laws to reduce energy demand.

In the second study (Chapter 3), we model and examine the energy-efficiency profile of individual houses in our sample. For this we use a load-disaggregation software, Princeton Scorekeeping Method (PRISM), which processes historical weather data and monthly utility usage data using an iterative regression approach to compute three energy-efficiency parameters: 1) baseload consumption for end-uses, which do not change with weather (e.g., lighting, refrigerator, and water heater), 2) heating slope, which is a function of the building's shell insulation and the efficiency of the heating (or cooling) unit, and 3) reference temperature, which is the outside temperature at which the house turns on heating (or cooling). These parameters further make up the normalized annual consumption (NAC), which is the weather-adjusted annual utility consumption for a typical year for given house. We then proceed to regress these parameters against the publically available data to study the extent we can extract statistical insight for residential energy-efficiency profiling using publically available information. The motivation behind this study is to build predictive models that can assess the different end-uses that constitute a house's energy load. This approach is meaningful for utility-driven, energy-efficiency targeting as two houses with the same annual utility usage may have different end-use-based energy-efficiency intervention potential. Further, this framework helps separate between engineering-driven (HVAC replacement) or behavioral intervention (adjusting thermostat settings) potential. Thus, utilities can predict changes in disaggregated load with respect to changes in demographic and structural characteristics at an individual-house level.

The third study (Chapter 4) uses the PRISM-computed energy-efficiency parameters to determine the savings potential in individual houses using different interventions both physical (e.g., new furnace, improved ceiling insulation, or replacing an old refrigerator) and behavioral (e.g., lowering the thermostat setting in the winter or raising it during the summer). To estimate the energy-efficiency reservoir for the houses in our sample, we model the reduction in energy consumption (and utility bills) that would occur if all the houses that were worse than the sample's median with regard to heating slope, thermostat setting, and baseload consumption could be improved to the median value. Because we have profiled each house in the sample, we know which houses could benefit from improvements to the physical system (e.g., new ceiling or wall insulation, or installation of a high-efficiency furnace) and keep their thermostats at setting that is "uncomfortable" relative to other houses in the sample (e.g., lower in the winter and higher in the summer). Following the approach taken in the first two papers, we regress the efficiency potential from different interventions against the publically available data to create a model that identifies houses with large savings potential for specific interventions. This approach enables utility program managers to predict the savings potential by house and by end-use using publically available data and help construct an EE reservoir map for targeted EE deployment. This in turn can allow utilities to allocate the right resources to specific houses using the right EE messaging and intervention to meet EE targets in an informed and analytical fashion. In addition, home owners, who were unaware of the savings potential, could be notified. Policymakers, who are designing rebate programs, could accurately forecast expected benefit from differently funded programs. Environmental groups

In the second half of this paper, we hypothesized that utility savings that would be realized from improvements to the physical system would be used to adjust the thermostat to a more comfortable settings. These houses would experience a “rebound effect” where some (or all) energy savings are used to improve comfort. To our knowledge, this is the first attempt at predicting rebound for the housing stock in a broad region. Because of our large dataset, we are able to build predictive models of which house are likely to experience rebound.

The reader should note that there are two potentially major limitations to this work. First, the underlying publically-available demographic information is from voter registration records. In other words, our house occupancy models do not account for children under 18 or unregistered adults. Only adults that have registered to vote are included. According to the Alachua County Supervisor of Elections,[1] approximately 80% of eligible voters are registered in the county. The majority of these eligible but unregistered voters is because of the large university student populations associate with the University of Florida and Santa Fe College. However, since our study focuses on single-family houses with multi-year, continuous utility data, the impact of these transient and “missing” occupants will be minimal. Determining the energy efficiency reservoir associated with students living in apartments near large universities would be a different study.

Second, the third and fourth chapters use PRISM, an iterative optimization model, to determine the disaggregated load variables. Because of the model’s design, it expects utility usage to follow a general trend (e.g., colder weather is matched with increasing heating bills). Missing data and trends that are counter to the model’s logic will result in a failed model fit. Approximately 15 percent of the houses in our dataset had utility-bill values that could not be modeled by PRISM. The characteristics of these houses and their occupants were not different than the houses for which valid PRISM models were found. So though the error rate of PRISM is troubling, we do not believe that it caused a significant bias in our analyses.

Having discussed the three studies we conclude with policy implications for our work and potential future research areas.

[1] Gainesville is located in Alachua County in Florida.

2. Statistical Modeling of Residential Utility Usage

Abstract

Energy efficiency (EE) and energy conservation today are recognized as the low-hanging fruit of energy-reduction strategies. However, the potential benefits of energy efficiency are often unrealized because of market failures and market barriers. The overarching objective behind this work is to merge publicly available data (e.g., property tax dataset for physical properties of houses and voter registration data set for demographic house properties), with utility consumption histories and extract statistically significant insight on energy efficiency and consumption for a group of houses (n=7,022) in Gainesville, FL. This study investigates the statistical descriptive power of publically available information for modeling utility usage. We first examine the deviations that arise from monthly utility usage reading dates as reading dates tend to shift and reading periods tend to vary across different months. Then we run regression models for individual months which in turn we compare to a yearly regression model which accounts for months as a dummy variable to understand whether a monthly model or a yearly model has a larger statistical power.

2.1. Introduction

Energy efficiency (EE) and energy conservation today are recognized as the low-hanging fruit of energy-reduction strategies (NAS, 2010). In recent years, several states have recognized the potential for energy efficiency to reduce energy consumption and pollutants' emissions, as well as possibly avoiding new generation construction. Thus, in order to promote energy efficiency, 24 states to-date have enacted Energy Efficiency Resource Standards (EERS) and set reduction targets for energy consumption (ACEEE, 2011). These targets have annual reductions goals that range between 0% and 2.2% from a baseline year (ACEEE, 2011).

To achieve these energy efficiency goals, several strategies can be pursued. One is the use of demand-side management (DSM) programs. Almost $7 billion was spent in rate-payer-funded DSM programs at a national level in 2011 and it is anticipated that a total of $12 billion will be spent by 2020 (IEE, 2012). Although a large number of states are meeting their demand reduction targets, with the coming more aggressive reduction goals, the exercise at hand will become more difficult. The potential benefits of energy efficiency are often unrealized because of market failures and market barriers. Some of these include information barriers, split incentives, hidden costs, transaction costs, high discount rates and heterogeneity among potential adopters (Jaffe and Stavins, 1994). Additionally, unpriced costs and benefits, misconstrued fiscal and regulatory policies, and insufficient and inaccurate information are recognized as market failures (NAS, 2010). Low priority of energy issues, incomplete markets for energy efficiency and limited access to capital (e.g., loans), further constitute market barriers for efficiency deployment (NAS, 2010).

Energy audits, one of the primary vehicles in promoting efficiency deployment at a building level, have been experiencing low penetration rates because they usually require consumers to seek audits independently and most consumers are unaware of need or value of an audit (Neme et al., 2011). Neme et al. estimate that state- and utility-sponsored audit programs reach less than two percent of homes each year (Neme et al., 2011). This is further exacerbated by low retrofit-project conversion rates followed by an energy audit, i.e., only a subset of audits evolve into a retrofit project because of high upfront costs and low return on investment (ROI) associated with the average consumer. The U.S. Department of Energy estimated that less than one percent of homes have had energy retrofits specifically to save energy (Lee, 2010). Other motivations may include health and safety reasons and comfort improvement. Palmer et al. (2011) surveyed approximately 500 energy auditors about the reasons why homeowners do not get audits. Responses in order of importance are: 1) lack of finances; 2) lack of knowledge on what an energy audit is; 3) lack of awareness on energy audits' existence; 4) high perceived or actual costs of audits. Palmer et al. (2011) further ranked the reasons for why homeowners make improvements (in decreasing frequency): 1) high savings; 2) low improvement cost; 3) non-energy benefits; 4) financing availability.

Despite the benefit, conducting energy audits for a majority of U.S. residential buildings is impractical because of the number of residential structures (120 million), high costs of audits ($200-$300/house), and limited workforce of qualified home energy performance professional.

Various strategies have been recommended to tackle the aforementioned barriers and enhance energy efficiency and conservation (Nowak et al., 2011). For example, Nowak et al. (2011) underscored the importance of identifying and prioritizing targeted technologies and end-uses. They emphasized that energy efficiency programs should prioritize their investments within consumer bases (i.e., focusing on potential high-savings potential projects first, before targeting a broader participation). Employment of innovative advertising and promotional channels would help enhance energy efficiency adoption but they need to be focused to achieve maximum impact. Similarly, Fuller et al. (2010) propose studying the population and finding and targeting early adopters.

Steemers and Yun (2009, 2011) studied the role of climate, occupant behavioral aspects and physical building characteristics in residential energy consumption. In their econometric study, the authors analyzed 4,822 housing units extracted from micro data from the US Department of Energy's Residential Energy Consumption Survey data from 2001, and found that, apart from climate, occupant behavioral aspects and socio-economic aspects (e.g., house income) are critical in terms of energy consumption, in particular for heating and cooling, through their effects on choices on physical building and appliance characteristics.

MacSleyne (2007) in her doctoral dissertation identified inefficient houses in Pittsburgh, PA, by using monthly natural gas consumption and physical and social characteristics of houses. Her work depended on building-level information – this granularity of data is highly uncommon because of utility-side proprietary issues or availability of data in a digitally appropriate format. She further explored ways to find cost-effective energy efficiency interventions for a house in order to reduce annual natural gas costs and/or improve indoor comfort. She also formulated strategies to prioritize low-income houses for a subsidized weatherization program (MacSleyne, 2007). This combination of datasets for 10 thousand houses is the largest number of individual houses studied at this level of detail to date. Given the extensive data at household level her work produced novel insight compared to the rest of the similar literature as discussed below.

Min et al. (2010) used Residential Energy Consumption Survey (RECS) data with U.S. census 2000 five-digit zip-code-level information and climate division-level temperature data to build a regression-based statistical framework to model space heating, cooling, water heating and appliance energy end uses, fuels used and carbon emissions at a zip-code-level resolution for the entire U.S. Min et al. (2010) acknowledges that the absence of high-resolution information on residential energy consumption has been and still is a significant impediment to the effective development and targeting of residential energy efficiency programs.

Jacobsen and Kotchen (2010) studied how a change in building code in Florida affected energy conservation in Gainesville, FL. They used building-level residential billing data for electricity and natural gas consumption. In their difference-in-differences analysis, which is an econometric technique that measures the effect of a treatment at a given period in time, they found that the building-code change, which took effect in 2002, resulted in a 4-percent decrease in electricity consumption and a 6-percent decrease in natural gas consumption. This build-code change involved improvements to baseline heating system, baseline air-distribution system and the Solar Heat Gain Coefficient.

Oikonomou et al. (2009) discuss how energy efficiency can be considered as resource reservoir extractable through actions and investments. This reservoir comprises both a private and a public economic value. The private value constitutes monetary return to utilities and consumers for energy efficiency investments, whereas the public value relates to externalities and public goods such as air pollution, energy security and continuity of service (Oikonomou et al., 2009). Much like other energy services, energy efficiency “reservoirs” are expected to vary significantly geographically.

Efficiency program design and implementation could significantly benefit from prioritization of efficiency projects and measures among consumer-bases through building a better understanding of the variation of the potential. This in turn can facilitate achieving higher consumption-reduction targets and a more cost-effective return for electric utilities. Identifying which buildings possess a greater savings potential may also help home energy-performance professionals overcome difficulties arising from low consumer conversion rates that are exacerbated by low return on investment for low savings potential consumers. Prioritizing efficiency projects can further improve job creation in the home energy performance industry.

This work explores the effects of modeling physical and social characteristics of houses in Gainesville, FL, on resource consumption (electricity, water and natural gas) and efficiency potential. Using publically available data on physical and social characteristics of houses can pave a path to understand and predict residential energy consumption and efficiency potential in a world where consumption information is highly private.

2.2. Data

The data were obtained from Program For Resource Efficient Communities (PREC) at University of Florida[2]. PREC collected the utility usage data from Gainesville Regional Utilities. The data include monthly utility consumption of single-family houses: 10,056 houses for electricity, 7,202 houses for natural gas and 10,166 for water, for years 2009, 2010 and 2011. The 7,202 houses with natural gas accounts use natural gas a primary heating source. Houses without natural gas usage were assumed to use electricity as primary heating source. To put this into context the U.S. Department of Energy's Residential Energy Consumption Survey (RECS) from 2009 contains data on 948 houses for the entire state of Florida.[3]

Not all of these houses have meter readings for all of the months between 2009 and 2011. For consistency, houses that do not have at least 33 or more readings for electricity, natural gas and water for 2009-2011 were dropped because a high number of missing readings for a given house could cause inaccuracies in our regression model. This resulted in 11 readings per year on average for a given house. This resulted in the final dataset size of 9,904 houses for electricity and water, and 7,096 houses for natural gas, since not all houses use natural gas and

[2] http://www.buildgreen.ufl.edu/

[3] http://www.eia.gov/consumption/residential/data/2009/

can use electricity instead. Further, houses that use a fuel other than electricity or natural gas for heating, e.g., heating oil or solar heater, were removed to avoid discrepancies. Thus, the final sample size for electricity, natural gas and water, was 9,461, 7,022 and 9,460, respectively (Table 1). The voter registration and property tax datasets have information for 18,190 houses and they were retrieved from Alachua County by PREC. A detailed description for each variable in these data sets is given in the Appendix at the end of this chapter. Since we have monthly consumption data for only a subset, we could only use the respective portion of the voter registration and property tax datasets. A comparison of the characteristics of the 9,461 houses in the electricity bill dataset and the remaining 12,289 houses in Gainesville showed minor statistical differences (Tables 2 and 3). Since this study covers only the 2009-2011 period and the corresponding utility consumption, the newest building in the sample was constructed in 2008.

Table 2. 1. Filtering constraints for the datasets showing number of houses

	Electricity	Water	Natural Gas
Total # of housing units	19,381	19,381	19,381
Total # of single-family units	10,056	10,166	7,202
≥33 & readings	9,904	9,904	7,096
Electric or natural gas heating	9,461	9,460	7,022
Final sample	9,461	9,460	7,022

Exploratory statistical work was conducted on the physical and demographic characteristics. Tables 2-5 show the percentile distribution of variables in the final sample for houses with electricity usage (n=9,461) and the rest of the population. The median house in the Gainesville sample was 35 years old, had $25,000 in land value and $97,000 in building value, whereas the median house outside the final sample was 36 years old and the corresponding land and building values were $25,000 and $85,000, respectively. Both groups had the same median in terms number of bedrooms, bathrooms, stories and number of occupants (3, 2, 1 and 2) (Tables 2-5). The quantified differences between Tables 2-5 are shown in Table 6 and 7. Overall, the final sample has higher building values, higher square footage and older occupants than the houses outside sample. Given the houses outside the final sample are largely multi-unit buildings and Gainesville is home to approximately 50,000 University of Florida students this comparison confirms our expectations.

Table 2. 2. Distribution of physical characteristics of houses in the final sample[4] of houses with electricity use (n=9,461)

	Percentile				
	5th	**25th**	**50th**	**75th**	**95th**
Land value ($1000)	10	18	25	34	50
Building value ($1000)	48	73	97	129	173
Misc. value ($1000)	0.2	0.7	1.6	3.2	7.6
Tax amount ($1000)	0.4	0.8	1.6	2.4	3.6
Heated area (1000 sqft)	0.9	1.2	1.5	2.0	2.5
Actual area (1000 sqft)	1.2	1.6	2.0	2.6	3.3
Age of the building	11	27	35	38	42
# of bedrooms	2	3	3	3	4
# of bathrooms	1	2	2	2	2.5
# of stories	1	1	1	1	1.5

Table 2. 3. Distribution of demographic characteristics of houses in the final sample[5] with electricity use (n=9,461)

	Percentile				
	5th	**25th**	**50th**	**75th**	**95th**
Avg age of occupants	18	32	44	66	80
Avg years of occupancy	3	8	14	24	34
# of occupants	1	1	2	2	3
# of teenagers	0	0	0	0	0
# of adults	0	0	1	2	3
# of seniors	0	0	0	1	2
# of republicans	0	0	0	1	2
# of democrats	0	0	1	2	2
# of males	0	0	1	1	2
# of females	0	1	1	1	2

[4] Please see appendix for the definitions of the variables.

[5] Please see appendix for the definitions of the variables. The information was extracted from voter registration database. # of teenagers only includes teenage occupants of voting age (18 and 19).

Table 2. 4. Distribution of physical characteristics of houses outside the final sample

	Percentile				
	5th	**25th**	**50th**	**75th**	**95th**
Land value ($1000)	8.8	18	25	35	50
Building value ($1000)	38	64	85	109	142
Misc. value ($1000)	0	0.5	1.1	2.1	4
Tax amount ($1000)	0.7	1.7	2.4	3.2	4.1
Heated area (1000 sqft)	0.8	1.1	1.4	1.7	2.1
Actual area (1000 sqft)	0.9	1.4	1.8	2.2	2.8
Age of the building	8	28	36	39	43
# of bedrooms	2	3	3	3	4
# of bathrooms	1	1	2	2	2
# of stories	1	1	1	1	1.5

Table 2. 5. Distribution of demographic characteristics of houses outside the final sample

	Percentile				
	5th	**25th**	**50th**	**75th**	**95th**
Avg age of occupants	22.9	26.8	32.7	42.5	55.8
Avg years of occupancy	1	4	6	11	17
# of occupants	1	1	2	2	3
# of teenagers	0	0	0	0	0
# of adults	0	1	2	2	3
# of seniors	0	0	0	0	1
# of epublicans	0	0	0	1	1
# of Democrats	0	0	1	1	2
# of males	0	0	1	1	2
# of females	0	0	1	1	2

Table 2. 6. The difference between the distributions of physical characteristics of houses in the final sample and outside the final sample

	Percentile				
	5th	**25th**	**50th**	**75th**	**95th**
Land value ($1000)	1.2	0	0	-1	0
Building value ($1000)	10	10	12	20	31
Misc. value ($1000)	0.2	0.2	0.5	1.1	3.6
Tax amount ($1000)	-0.4	-0.8	-0.8	-0.7	-0.6
Heated area (1000 sqft)	0.1	0.1	0.2	0.3	0.4
Actual area (1000 sqft)	0.2	0.2	0.3	0.4	0.5
Age of the building	3	-1	-1	-1	-1
# of bedrooms	0	0	0	0	0
# of bathrooms	0	1	0	0	0.5
# of stories	0	0	0	0	0

Table 2. 7. The difference between the distributions of demographic characteristics of houses in the final sample and outside the final sample

	Percentile				
	5th	**25th**	**50th**	**75th**	**90th**
Avg age of occupants	-4.9	5.2	11.3	23.5	24.2
Avg years of occupancy	2	4	8	13	17
# of occupants	0	0	0	0	0
# of teenagers	0	0	0	0	0
# of adults	0	-1	-1	0	0
# of seniors	0	0	0	1	1
# of republicans	0	0	0	0	1
# of democrats	0	0	0	1	0
# of males	0	0	0	0	0
# of females	0	1	0	0	0

Similarly, the distributions in utility consumptions by month were investigated. It is common that utilities collect monthly consumption data in irregular frequencies. In other words, the meter reading dates can vary from month to month resulting in non-uniform billing periods. If the analyst assigns consumption values only to months in which the reading was measured and does not attribute the consumption between consecutive months accounting for the billing periods and the number of days in billing periods, discrepancies can emerge. We compared two approaches: 1) monthly utility usage is directly attributed to the month the reading was measured in; 2) monthly utility usage is adjusted via distributing the usage amount between consecutive months based on the reading date and number of days in the billing period. To clarify, for the second approach, we allocated adjusted consumptions based on monthly readings and readings date proportionally across consecutive months. For example, if a meter reading was conducted on March 16 and the billing duration was 26 days, then the adjusted March consumption would contain $\frac{16}{26}$ of the reading amount where the remainder, $\frac{10}{26}$, would be attributed to the adjusted February consumption. The second half of March would come from the next billing period. Although both unadjusted and adjusted consumptions show seasonal effects (e.g., higher electricity consumption in summer months due to cooling), there are significant monthly differences in the two approaches where the first one does not adjust for billing lag and the second one does.

Figures 1-3 show the variation of the average (adjusted? How?) daily electricity, natural gas and water consumption by month for 2009-2011. As expected, summer months experience the most electricity consumption in Florida because of increased air conditioning. Conversely, the least natural gas consumption occurs in summer months because of decreased heating-related consumption. Monthly water consumption is relatively flat with a slight increase occurring in summer months. To put monthly utility consumption values into context, in 2010 the average American house consumed 11,496 kWh (31.5kWh/day, 958kWh/month) and 980 therms of natural gas (2.7therms/day, 81.7therms/month) (EIA, 2012). Further, the average American house consumes 120 thousand gallons of water (including outdoor usage) annually (300gal/day, 10,000gal/month) (America Water Works Association, 2012). For further context, RECS data suggest that the average Florida household consumes 15,000kWh/year[6].

Interestingly, billing lag and lack of adjustment thereof can result in substantial differences in transition months (Figures 1-4) where the corresponding preceding/succeeding months experience a change in temperature that influences utility consumption. These transition months were May and September for electricity; March and November for natural gas; and March and June for water. Depending on the task at hand the aforementioned adjustment may

[6] http://www.eia.gov/consumption/residential/data/2009/

become critical. As such we have accounted for billing lag in our regression work that is explained in the following section.

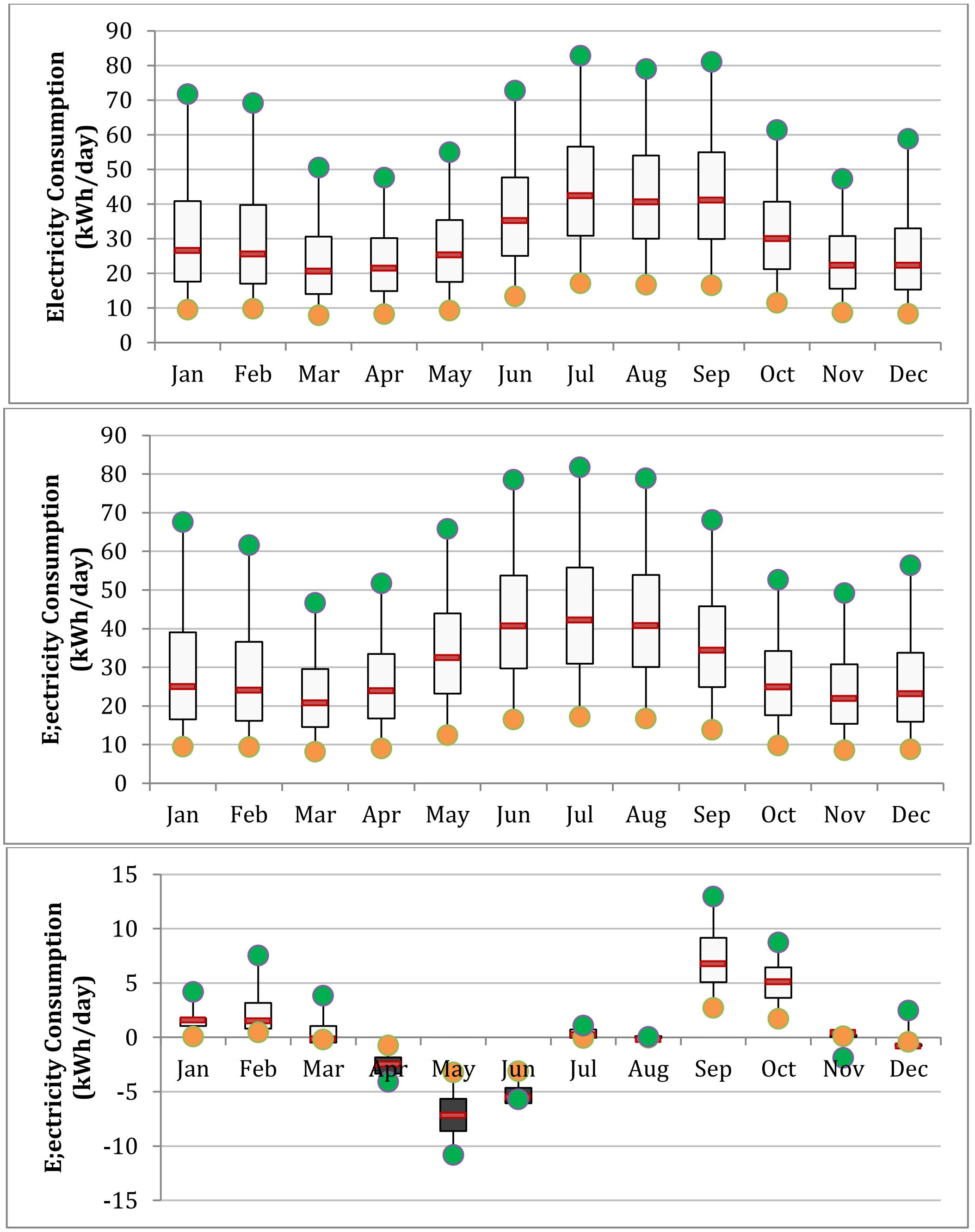

Figure 2. 1. 5th, 25th, 50th, 75th and 95th percentile distribution of average daily electricity consumption by month (2009-2011). The first figure shows consumption unadjusted for billing period, the second one adjusted for billing period, the third one is the difference between the two.

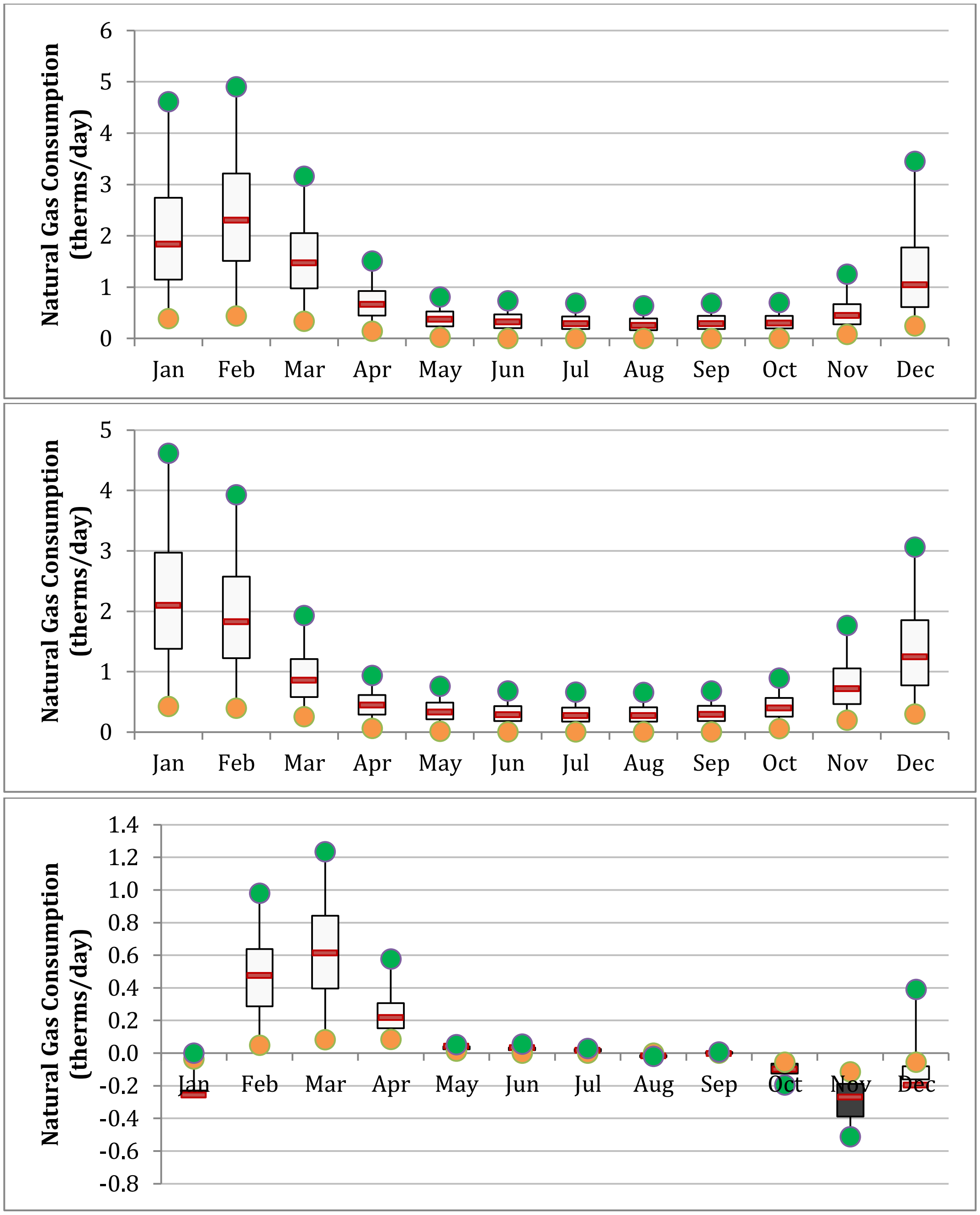

Figure 2. 2. 5th, 25th, 50th, 75th and 95th percentile distribution of average daily natural gas consumption by month (2009-2011). The first figure shows consumption unadjusted for billing period, the second one adjusted for billing period, the third one is the difference between the two. This sample includes all houses that use natural gas (n=7,022).

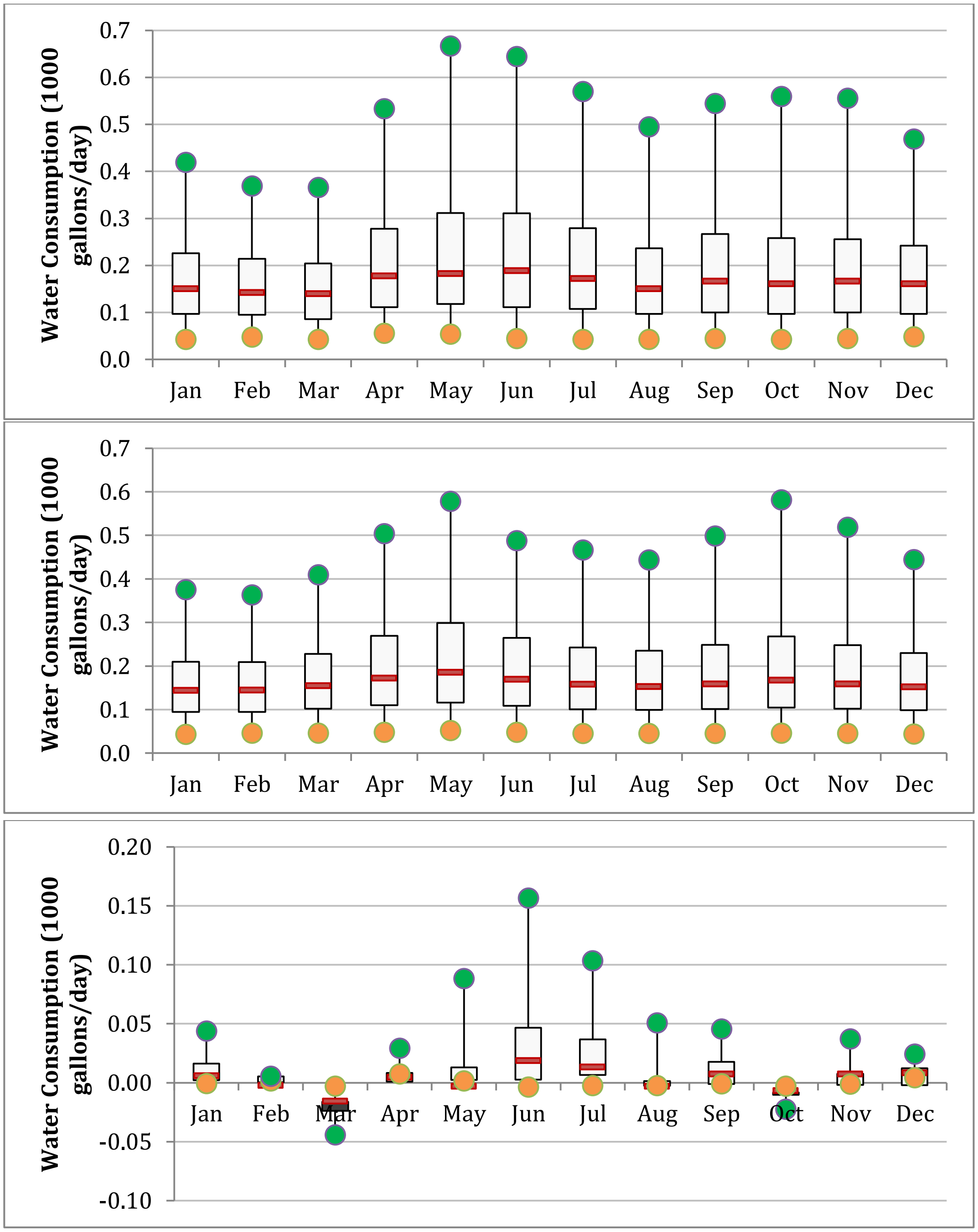

Figure 2. 3. 5th, 25th, 50th, 75th and 95th percentile distribution of average daily water consumption by month (2009-2011). The first figure shows consumption unadjusted for billing period, the second one adjusted for billing period, the third one is the difference between the two.

Electricity and natural gas consumptions, in kWh and therm, respectively, were aggregated in btu terms to construct total site-delivered energy consumption[7]. Electricity component in energy consumption increases in summer months due to increased air conditioning (Figure 4).

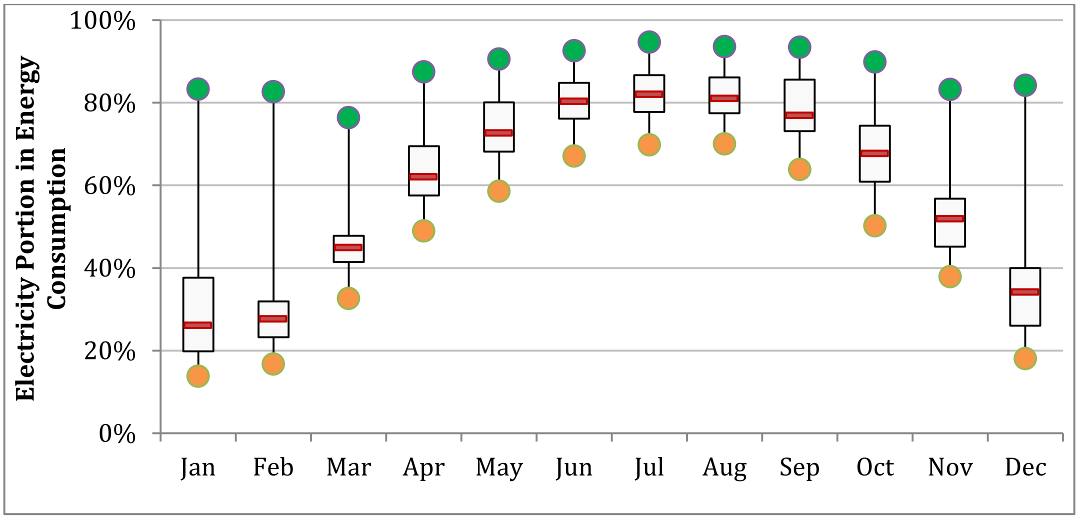

Figure 2. 4. 5th, 25th, 50th, 75th and 95th percentile distribution of percentage of electricity consumption in total energy consumption.

Temperatures for the study period (2009-2001) were not uncharacteristic for the Gainesville region. Figure 5 shows the daily maximum and minimum for 2000-2012 as obtained from the U.S. National Oceanographic and Atmospheric Administration.

[7] 1 kWh = 3,412 btu; 1 therm = 100,000 btu

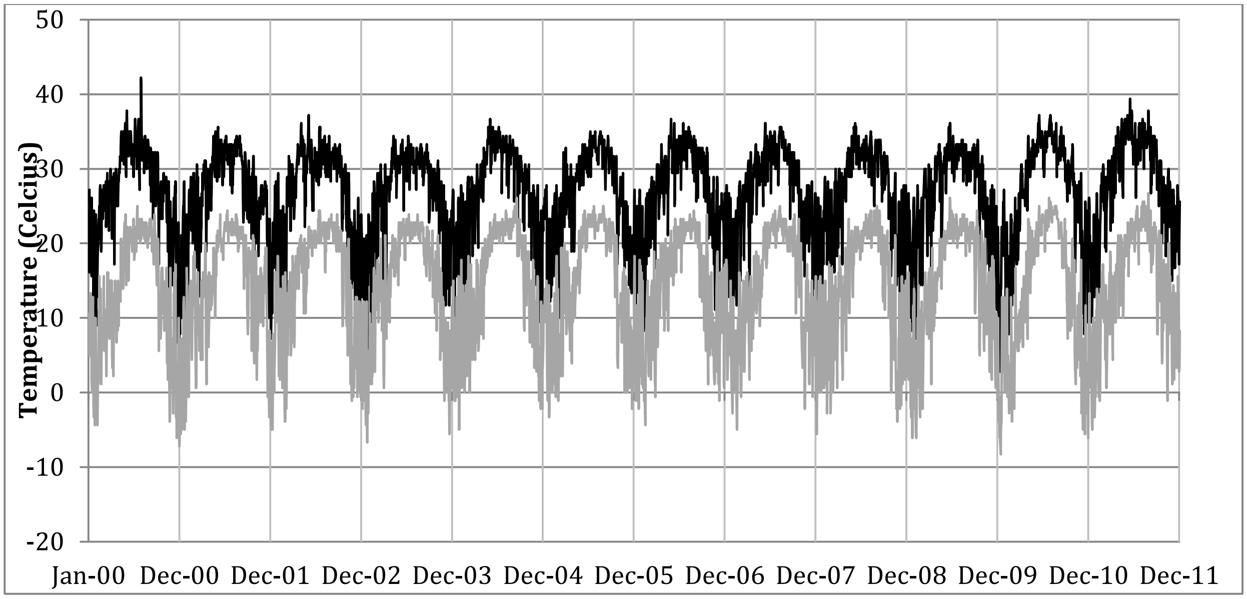

Figure 2. 5. Daily maximum and minimum temperatures between 2000 and 2012.

2.3. Models

The overarching objective of this study is to investigate how publically available data (i.e., property tax information and voter registration records) can be used to model residential utility usage using two regression models. The first models each individual month separately the second uses a yearly model with dummy variables for each month. For both approaches, stepwise regressions were run[8]. An additional regression was conducted for total energy that integrates electricity and natural gas use in btu terms. This technique is to study the explanatory power of the proposed variables on different months.

This results of this exercise is helpful for utilities for two reasons: 1) Utilities can use our models to predict monthly changes in demand as its driven by changes in the structural and demographic characteristics in their service territory; 2) Predicting utility usage can be translated into energy-use intensity per squarefoot which can be used as a first-cut metric for energy-efficiency targeting in their service territory to meet their state-mandated demand-reduction targets.

[8] Stata created by StataCorp was used to run the regression models.

Months for different years were combined (i.e., January 2009, 2010 and 2011 were combined). In other words each observation in the regression models uses 36 months utility usage information on average covering 2009, 2010 and 2011. December data only included 2009 and 2010 because billing data for January 2012 was not available to complete the month.

2.3.1. Individual Monthly Regressions

For each utility use (i.e., electricity (kWh/day), natural gas (millitherms/day), water (gallons/day), energy (btu/day), electricity only (kWh/day)) twelve separate stepwise regressions were run, one for each month, resulting in 60 regressions. Electricity only curve is for houses that do not use natural gas and use electricity for primary heating source. Average daily utility usage for each month was regressed against the physical and demographic characteristics as described in the previous section. Figure 6 shows that the R^2 varies between 0.32 and 0.39 for electricity; 0.25-0.32 for electricity for houses with no natural gas consumption; 0.06 and 0.20 for natural gas; 0.08 and 0.14 for water; and 0.33 and 0.42 for energy. Given this exercise involves incorporating demographic factors in utility usage, the resulting R^2 values can be considered as relatively high.

The statistical explanatory power of the Total Energy, Electricity and Natural Gas models are the highest during winter months and relatively flat during the rest of the year, perhaps because of increased heating during winter months. However, the Natural Gas model's R^2 is substantially lower than that of the Total Energy and Electricity models. On the contrary, Electricity Only model has the highest explanatory power during summer months, perhaps because of increased cooling. The Water model's performance is relatively flat throughout the year.

These fluctuating patterns may be tied to the influence of weather on residential utility use, e.g. summer (winter) months experience higher energy use because of increased cooling (heating).

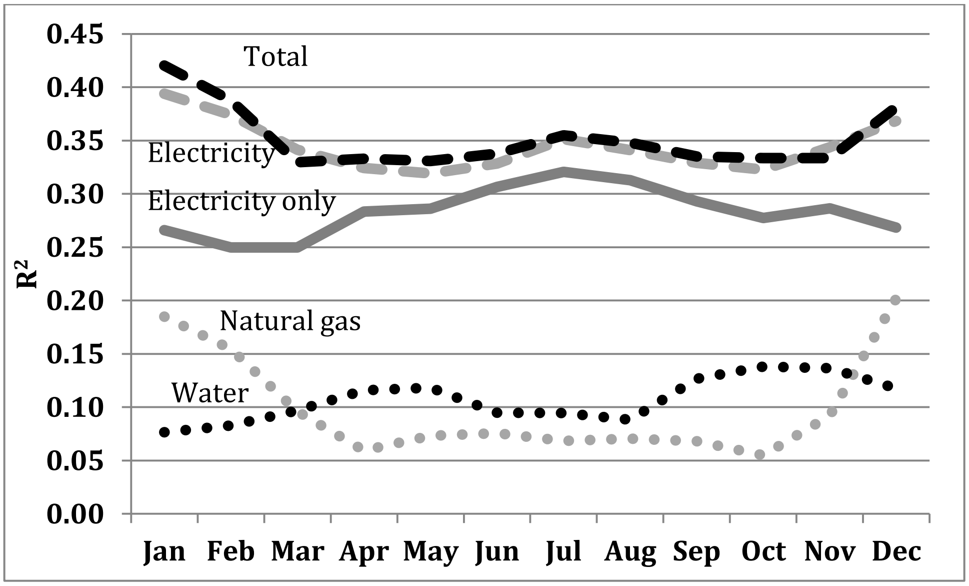

Figure 2. 6. R^2 by month and utility. Electricity only curve denotes houses that do not use any natural gas.

Electricity, natural gas and energy consumption share a comparable theme on independent variables' influence on utility usage (Tables 8-12). Most independent variables are statistically significant across different months. Some findings do not oppose common sense, e.g. higher property value and higher square footage implies more consumption. Heated area has a larger impact on all utility consumption than actual area. Further, heated area's influence, which can be considered as conditioned area, rises substantially during summer months which can be tied to increased cooling loads. Houses with electric heating (Fuel dummy=1) consume more electricity than homes with natural gas heating. This effect is more significant in winter months. Seniors tend to consume less than teenagers on average. Republicans consume more electricity than Democrats during winter months and less during summer months. Older buildings and buildings with higher average age of occupants tend to have higher consumption. Similar regression coefficients were found for houses that do not consume any natural gas (Table 9). Houses that use electric heating (Fuel dummy=1) consume less natural gas (Table 10).

Similar to electricity consumption, higher property values and square footage result in higher natural gas consumption where summer months experience less consumption than winter months. Older buildings lead and higher average age of occupants lead to higher consumption.

From a statistics standpoint the coefficients' magnitude and signs should be interpreted cautiously and not independently because of the uncertainty in independent variables and all of the other potential variables, e.g., number of children, efficiency of the HVAC unit, that were not included in the models. It should be underscored that the objective of this exercise was to demonstrate the statistical explanatory power of the publically available information in predicting utility usage and the coefficients are likely to change for similar models applied to other geographic regions.

Table 2. 8. Average daily electricity consumption (kWh/day) regressions by month for houses that use both natural gas and electricity. The values indicate the beta coefficients for each regression which are statistically significant at p≤0.10.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Property value ($1,000)	0.07	0.06	0.06	0.06	0.07	0.07	0.05	0.06	0.06	0.04	0.05	0.07
Property tax amount ($1,000)	2.18	2.08	1.56	1.47	1.29	1.15	1.37	1.46	1.68	1.57	1.65	2.05
Total property value ($1,000)	-0.07	-0.06	-0.05	-0.04	-0.06	-0.06	-0.05	-0.05	-0.05	-0.03	-0.04	-0.07
Actual area (1,000 sqft)	3.16	3.16	2.65	3.02	3.34	3.64	4.01	3.11	2.66	2.68	2.53	2.90
Heated area (1,000 sqft)	5.89	4.29	3.50	3.94	6.80	9.10	9.57	9.57	7.69	4.56	4.05	6.24
Age of the building	0.28	0.25	0.15	0.12	0.16	0.21	0.22	0.20	0.16	0.13	0.17	0.26
# of bedrooms	-	-	-	-	-	0.49	0.51	0.69	0.47	-	-	-
# of bathrooms	-0.85	-	0.34	0.96	0.77	0.74	1.30	0.95	0.74	0.90	0.61	-0.76
# of stories	-0.46	-0.45	-0.51	-0.62	-0.96	-1.14	-1.23	-1.01	-0.80	-0.55	-0.33	-
Fuel dummy	2.36	1.90	0.95	0.42	0.43	-	0.44	0.48	0.37	0.45	1.15	2.13
Natural gas acc. dummy	-21.24	-17.74	-8.89	-4.36	-2.55	-1.36	-0.91	-1.19	-2.12	-4.06	-7.73	-18.09
Avg. age of occupants	-	-	-0.02	-0.05	-0.05	-0.04	-0.04	-0.06	-0.06	-0.04	-	-
Avg. year of occupancy	0.02	0.03	0.04	0.06	0.04	0.04	0.04	0.03	0.04	0.04	0.04	0.05
# of teenagers	1.98	2.16	1.67	1.64	2.48	2.87	2.41	2.04	1.89	1.45	1.86	2.32
# of seniors	-0.94	-0.95	-0.87	-1.04	-1.68	-1.96	-1.87	-1.87	-1.53	-1.16	-1.17	-0.91
# of occupants	2.85	2.50	2.53	3.63	3.97	4.05	4.20	4.69	4.30	3.40	2.21	2.55
# of republicans	0.70	0.66	1.08	1.48	1.94	2.02	2.17	2.11	1.98	1.50	1.31	1.07
# of democrats	0.88	0.53	0.27	-	0.56	0.69	0.59	0.67	0.55	-	0.41	1.17
# of females	-	-	-	-0.26	-	0.42	0.46	-	-	-	-	-
# of males	-0.42	-	-	-	-	-	-	-	-	-	-	-0.47
Intercept	12.48	9.91	5.26	2.91	3.36	3.15	1.02	2.52	3.48	2.87	2.41	8.33
R^2	0.37	0.34	0.32	0.32	0.33	0.35	0.34	0.33	0.32	0.34	0.37	0.37
n	28,383	28,383	28,383	28,383	28,383	28,383	28,383	28,383	28,383	28,383	28,383	18,922

Table 2. 9. Average daily electricity consumption (kWh/day) regressions by month for houses that do not use natural gas. The values indicate the beta coefficients for each regression which are statistically significant at p≤0.10.

	Jan	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
Property value ($1,000)	0.07	0.07	0.05	0.04	0.04	0.03	0.03	0.03	0.04	0.04	0.02	0.08
Property tax amount ($1,000)	2.36	2.12	1.48	1.01	0.70	0.84	1.33	1.19	1.24	1.32	1.74	2.56
Total property value ($1,000)	-0.07	-0.05	-0.03	-	-	-	-	-	-	-	-	-0.08
Actual area (1,000 sqft)	-	-	-	-	-	-	-	-1.90	-2.18	-1.29	-	-
Heated area (1,000 sqft)	11.25	7.95	5.53	4.98	7.79	10.86	10.89	12.35	10.35	6.22	4.73	11.50
Age of the building	0.25	0.23	0.10	0.04	0.06	0.07	0.09	0.09	0.06	0.05	0.12	0.24
# of bedrooms	1.84	1.55	1.02	1.00	1.55	2.18	1.97	2.04	1.57	0.98	0.86	1.69
# of bathrooms	-1.05	-	-	-	-	-0.99	-	-	-	-	0.66	-1.48
# of stories	-2.79	-2.29	-1.23	-0.70	-1.33	-1.74	-1.75	-1.69	-1.43	-0.67	-0.72	-1.82
Fuel dummy	5.59	4.16	2.11	1.01	1.10	1.26	1.40	1.26	1.06	1.17	2.12	5.31
Natural gas account dummy	-	-	-	-	-	-	-	-	-	-	-	-
Avg. age of occupants	0.12	0.07	-	-0.04	-0.04	-	-	-0.05	-0.05	-	-	-
Avg. year of occupancy	-	-	0.04	-	-	-	-	-	-	-	0.04	0.08
# of teenagers	-	-	-	-	-	-	-	-	-	-	-	-
# of seniors	-2.16	-1.57	-1.23	-1.14	-1.87	-2.78	-2.52	-2.31	-1.83	-1.68	-1.19	-
# of occupants	-	1.65	2.36	3.44	4.03	3.40	3.31	4.24	4.08	2.71	2.21	2.46
# of republicans	1.76	1.59	1.69	2.20	2.48	2.51	2.71	2.66	2.50	2.26	1.82	1.83
# of democrats	2.69	1.88	1.37	1.12	1.67	1.79	1.70	1.80	1.60	1.25	1.55	3.24
# of females	1.24	-	-	-	-	0.86	1.02	0.87	0.58	-	-	-
# of males	-	-	-	-	-	-	-	-	-	-	-	-0.96
Intercept	3.25	2.39	3.56	4.22	4.75	5.73	4.45	5.16	6.10	4.68	1.24	-0.55
R^2	0.27	0.25	0.25	0.28	0.29	0.31	0.32	0.31	0.29	0.28	0.29	0.27
n	7,314	7,314	7,314	7,314	7,314	7,314	7,314	7,314	7,314	7,314	7,314	4,876

Table 2. 10. Average daily natural gas consumption (millitherms/day) regressions by month. The values indicate the beta coefficients for each regression which are statistically significant at $p \leq 0.10$.

	Jan	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
Property value ($1,000)	2.4	2.5	2.5	2.0	2.1	1.8	1.8	1.7	1.6	1.9	0.7	3.5
Property tax ($1,000)	-133.7	-77.3	-19.8	13.7	12.3	8.9	8.0	9.5	11.2	9.4	-	-115.7
Total property ($1,000)	1.9	-	-1.0	-1.4	-1.5	-1.4	-1.4	-1.3	-1.2	-1.4	-	1.8
Actual area (1,000 sqft)	117.2	154.0	49.4	17.2	22.5	27.6	21.9	19.1	15.5	30.1	57.4	-
Heated area (1,000 sqft)	566.6	443.4	156.5	-	-39.7	-39.0	-36.4	-34.1	-22.5	-	143.3	610.1
Age of the building	23.6	19.3	7.9	-	-1.1	-1.1	-0.9	-1.1	-1.0	1.4	7.3	28.7
# of bedrooms	-	-	-	13.7	17.1	10.2	8.9	12.5	10.7	8.1	-	-
# of bathrooms	-55.4	-	-	15.6	10.2	9.8	11.9	9.6	9.6	-	-	-100.0
# of stories	-163.3	-183.3	-77.5	-31.5	-12.6	-	-6.7	-	-8.2	-10.3	-46.8	-94.8
Fuel dummy	-122.1	-112.0	-28.9	-23.0	-21.1	-20.5	-17.3	-18.0	-20.7	-22.4	-33.0	-
Natural gas account dummy	-	-	-	-	-	-	-	-	-	-	-	-
Avg. age of occupants	11.0	10.9	6.1	1.9	-	-	-0.4	-0.6	-	1.6	5.9	10.0
Avg. year of occupancy	-	-	-	-	0.6	0.4	0.5	0.6	0.5	0.7	1.4	3.1
# of teenagers	294.8	255.8	126.2	65.0	43.3	31.0	28.6	28.2	42.8	69.1	155.5	308.9
# of seniors	-	-44.1	-36.2	-34.3	-30.1	-25.4	-17.7	-19.5	-29.2	-31.0	-36.4	-
# of occupants	-120.5	-192.5	-49.0	-	59.4	55.2	62.1	50.6	41.4	-	-96.1	-108.3
# of republicans	-38.9	-34.3	-18.5	-	-	-	-	-	-	-	-19.4	-42.0
# of democrats	62.4	38.5	28.3	11.5	6.4	7.4	6.8	5.6	5.8	14.0	19.0	87.9
# of females	-96.5	-	-47.1	-	-15.0	-13.6	-15.1	-	-	-	-	-127.2
# of males	-135.7	-34.9	-61.9	-	-22.0	-22.4	-24.5	-9.2	-9.8	-11.1	-20.9	-165.2
Intercept	186.3	221.9	200.4	264.6	236.4	216.5	217.8	208.9	218.7	202.9	91.8	-357.7
R^2	0.18	0.15	0.10	0.06	0.07	0.08	0.07	0.07	0.07	0.06	0.09	0.20
n	21,066	21,066	21,066	21,066	21,066	21,066	21,066	21,066	21,066	21,066	21,066	14,044

Table 2. 11. Average daily energy consumption (btu/day) regressions by month. The values indicate the beta coefficients for each regression which are statistically significant at p≤0.10.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Property value ($1,000)	0.6	0.5	0.5	0.4	0.4	0.4	0.3	0.3	0.3	0.3	0.3	0.7
Property tax ($1,000)	-2.6	-	4.3	6.5	5.7	4.8	5.5	5.9	6.8	6.3	5.2	-
Total property ($1,000)	-0.2	-0.2	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.2	-0.2	-0.2
Actual area (1,000 sqft)	10.1	14.9	9.7	11.9	13.2	14.7	15.6	12.3	10.5	10.8	10.2	-
Heated area (1,000 sqft)	67.5	49.6	24.7	12.3	19.4	27.7	29.6	30.0	24.2	16.2	26.1	72.7
Age of the building	2.6	2.2	1.1	0.4	0.5	0.6	0.7	0.6	0.5	0.5	1.1	3.0
# of bedrooms	-2.8	-	-	-	2.1	2.3	2.3	3.2	2.3	-	-	-3.5
# of bathrooms	-7.6	-	-	4.2	2.6	2.8	4.9	3.5	2.8	2.7	-	-9.7
# of stories	-11.6	-12.9	-6.4	-4.1	-4.1	-4.5	-4.7	-3.9	-3.3	-2.5	-3.8	-6.1
Fuel dummy	-	-	-	-	-	-	-	-	-	-	-	7.2
Natural gas account dummy	144.9	127.2	60.3	30.6	25.7	26.1	25.7	24.8	23.4	28.0	51.6	121.1
Avg. age of occupants	0.9	0.8	0.3	-	-0.2	-0.2	-0.2	-0.2	-0.2	-	0.4	0.8
Avg. year of occupancy	-	-	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.4
# of teenagers	31.2	28.2	16.1	11.3	12.4	12.6	10.8	9.7	10.2	10.9	18.7	33.5
# of seniors	-4.3	-5.6	-5.3	-6.7	-8.1	-8.4	-7.8	-7.9	-7.2	-6.5	-6.1	-4.3
# of occupants	-	-	6.4	11.4	16.4	16.7	17.3	19.3	17.7	10.6	-	-
# of republicans	-	-	2.3	5.4	7.0	7.3	7.7	7.6	7.1	5.5	3.4	-
# of democrats	8.4	5.1	3.1	1.2	2.6	3.1	2.6	2.9	2.5	1.7	3.2	10.8
# of females	-8.1	-5.3	-4.9	-	-	1.9	2.1	-	-	-	-	-10.6
# of males	-11.9	-7.9	-5.5	-	-	-	-	-	-	-	-	-14.6
Intercept	-91.2	-83.1	-30.2	-2.0	3.1	5.6	-0.9	4.1	6.8	-3.6	-37.1	-119.4
R^2	0.42	0.39	0.33	0.33	0.33	0.34	0.35	0.35	0.34	0.33	0.33	0.38
n	28,383	28,383	28,383	28,383	28,383	28,383	28,383	28,383	28,383	28,383	28,383	18,922

Table 2. 12. Average daily water consumption (gallons/day) regressions by month. The values indicate the beta coefficients for each regression which are statistically significant at $p \leq 0.10$.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Property value ($1,000)	0.5	0.2	0.1	0.3	0.4	0.2	0.2	0.2	0.2	0.4	0.3	0.5
Property tax ($1,000)	11.6	12.1	10.5	12.0	10.7	8.6	14.0	14.9	15.3	17.0	17.0	13.2
Total property ($1,000)	-0.4	-0.2	-	-	-	-	-	-	-	-	-	-0.3
Actual area (1,000 sqft)	-	-	-	-	11.8	-	-	-7.8	-	-	-7.8	-10.4
Heated area (1,000 sqft)	16.9	19.7	25.6	22.8	23.5	33.7	22.3	27.9	25.1	24.0	32.4	31.1
Age of the building	0.6	0.4	-	-0.5	-0.6	-0.6	-0.3	-	-0.5	-0.4	-	0.4
# of bedrooms	10.3	8.1	8.4	12.5	12.1	12.5	12.5	11.2	13.7	15.3	11.5	11.1
# of bathrooms	13.3	13.5	8.6	9.3	7.3	-	6.9	8.0	8.0	7.4	12.8	15.7
# of stories	-8.7	-9.5	-10.6	-19.9	-25.6	-21.2	-15.4	-16.8	-21.6	-26.4	-22.4	-15.7
Fuel dummy	-4.2	-	-	-4.3	-5.0	-4.2	-4.3	-5.5	-	-	-	-4.5
Natural gas account dummy	5.1	8.3	10.7	18.0	19.8	16.4	13.7	13.6	16.3	16.9	14.9	7.5
Avg. age of occupants	0.2	0.2	0.4	0.7	0.9	0.8	0.6	0.7	0.6	0.7	0.7	0.5
Avg. year of occupancy	0.3	-	-	0.3	0.4	-	-	-	0.3	0.4	0.4	0.3
# of teenagers	23.9	24.8	17.8	21.2	20.3	30.2	22.7	24.0	27.4	29.5	30.4	24.9
# of seniors	-11.8	-10.2	-7.0	-7.8	-9.8	-10.5	-8.6	-11.3	-8.2	-	-6.2	-11.1
# of occupants	16.1	20.1	15.1	6.0	-	-	6.0	7.2	6.2	-	-	9.0
# of republicans	-	-	3.5	8.2	11.7	7.7	7.5	5.4	10.0	11.1	7.6	4.5
# of democrats	-	-	-	-	-	-	-	-	-	-	-	-
# of females	5.5	-	3.4	7.1	9.4	13.9	9.4	4.6	3.6	7.6	6.2	7.2
# of males	-	-	-	-	-	5.4	-	-	-	-	-	-
Intercept	6.5	19.4	21.2	31.3	31.3	47.9	29.4	25.3	22.7	13.4	2.1	4.8
R^2	0.08	0.08	0.10	0.12	0.12	0.09	0.09	0.09	0.13	0.14	0.14	0.12
n	28,380	28,380	28,380	28,380	28,380	28,380	28,380	28,380	28,380	28,380	28,380	18,920

2.3.2. Regression for the Yearly Model with Dummy Variables for Months

In this approach, a single yearly model (with dummy variables for months) was used to predict utility consumption. All independent variables used in the first regression approach described in 3.1. were used in this approach as well as two additional independent variables: historical average monthly temperature and rainfall, to capture the impact of month-to-month change in weather.

As implied by the larger R^2 values, this model has a larger explanatory power for electricity, natural gas and energy consumption (Table 13). This can be attributed to the addition of the temperature and rainfall variables. The beta coefficients are comparable to those from the individual monthly regressions in terms of magnitude and sign (Table 13). Although greater R^2s suggest better explanatory power for this approach, predicting individual months' utility usage using this model can pose different error terms, i.e., the difference between predicted value and actual value, that are discussed in the next section.

Table 2. 13. Regression results for the yearly model. The values indicate the beta coefficients which are statistically significant at p≤0.10.

	Electricity	**Natural gas**	**Water**	**Total Energy**	**Electricity only**
Property value ($1,000)	0.06	1.99	0.3	0.4	0.05
Property tax amount ($1,000)	1.62	-20.42	13.4	4.2	1.54
Total property value ($1,000)	-0.05	-0.61	-0.1	-0.2	-0.02
Actual area (1,000 sqft)	3.07	47.95	-	11.6	-0.89
Heated area (1,000 sqft)	6.22	126.86	24.7	32.2	9.04
Age of the building	0.19	6.29	-0.1	1.1	0.11
# of bedrooms	0.28	11.97	11.5	0.6	1.53
# of bathrooms	0.46	-6.83	9.2	0.7	-0.31
# of stories	-0.69	-52.70	-17.8	-5.6	-1.51
Fuel dummy	0.91	-38.06	-4.0	-	2.19
Natural gas account dummy	-7.22	-	13.3	55.5	-
Avg. age of occupants	-0.03	3.75	0.6	0.2	-
Avg. year of occupancy	0.04	0.85	0.2	0.2	0.03
# of teenagers	2.06	116.16	24.7	16.7	0.81
# of seniors	-1.37	-30.05	-8.6	-6.6	-1.82
# of occupants	3.32	-	7.6	9.6	2.75
# of republicans	1.54	-10.64	6.6	4.6	2.10
# of democrats	0.54	23.29	-	3.8	1.66
# of females	0.10	-35.79	6.1	-1.4	-
# of males	-	-51.63	-	-2.7	-
Historical monthly avg. temperature	0.32	-89.10	2.8	-5.6	-0.41
Historical monthly avg. rainfall	-0.57	47.78	-7.3	9.1	1.80
Feb dummy	-3.34	51.25	-12.4	-9.1	-3.65
Mar dummy	-10.05	-573.26	-2.8	-80.1	-14.27
Apr dummy	-8.86	-520.60	-	-61.9	-10.05
May dummy	-	-	-	-	-
Jun dummy	6.84	-	-	-	-
Jul dummy	8.45	122.11	-24.1	19.9	3.64
Aug dummy	6.47	19.72	-25.6	7.5	1.81
Sep dummy	-	-135.70	-18.5	-12.6	-1.51
Oct dummy	-8.14	-482.89	5.2	-56.8	-9.01
Nov dummy	-8.70	-681.94	2.5	-70.7	-10.38
Dec dummy	-1.79	-162.32	-	-13.3	-1.24
Intercept	-17.33	7323.89	-167.6	414.2	33.49
R^2	0.40	0.50	0.11	0.42	0.34
n	331,135	85,330	331,135	245,770	331,100

2.3.3. Comparison between the Two Approaches

The second model's residuals, i.e., the difference between the predicted value and the actual value, for individual months were used to compute R^2 values for individual months with the following formula:

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}}$$

Where:

$$SS_{err} = \sum_{i}^{n} (actual\ value_i - predicted\ value_i)^2$$

$$SS_{tot} = \sum_{i}^{n} (actual\ value_i - mean\ of\ actual\ values)^2$$

To clarify the sum of squares were computed for each month, i.e., each model, was computed separately and the mean of actual values is calculated for the corresponding month.

As expected, the predictive power of the second model varies from month to month (Figures 7-11). Some months are explained better by the first approach (monthly model) than the second (Figures 7-11). The yearly model performs better for Total Energy, Electricity Only and Electricity for most of the months. However, the yearly model outperforms the monthly substantially for Natural Gas, probably given its significant dependence on outside temperature. No noteworthy differences were observed between the two approaches for Water.

Depending on the analyst's objective, she could be better off using the first approach, i.e., modeling individual months separately, if the goal is to model specific months. For a less specific purpose, a yearly model with dummy variables for months would be more appropriate.

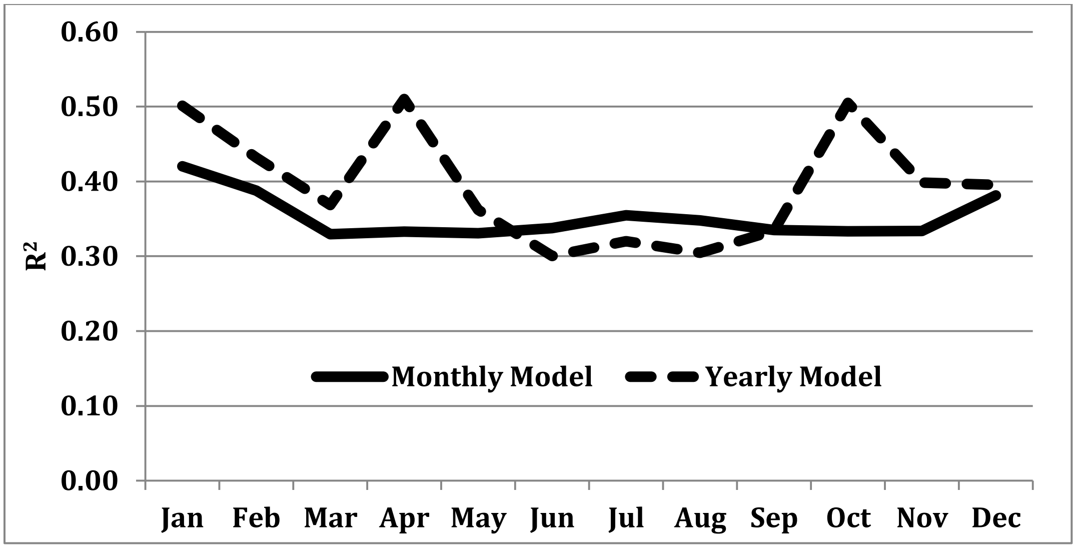

Figure 2. 7. R^2 values of monthly versus yearly model for Total Energy.

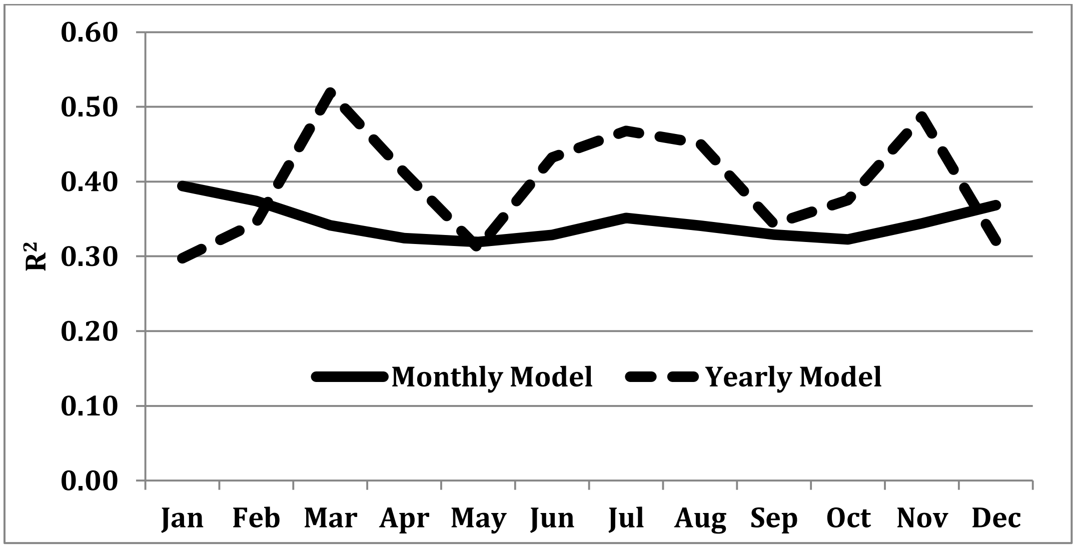

Figure 2. 8. R^2 values of monthly versus yearly model for Electricity.

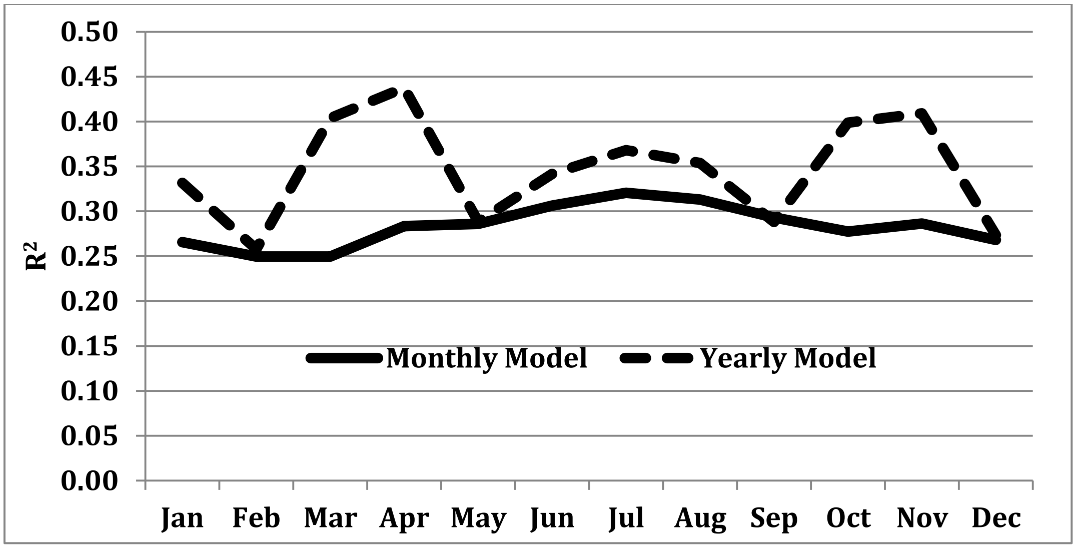

Figure 2. 9. R^2 values of monthly versus yearly model for Electricity Only.

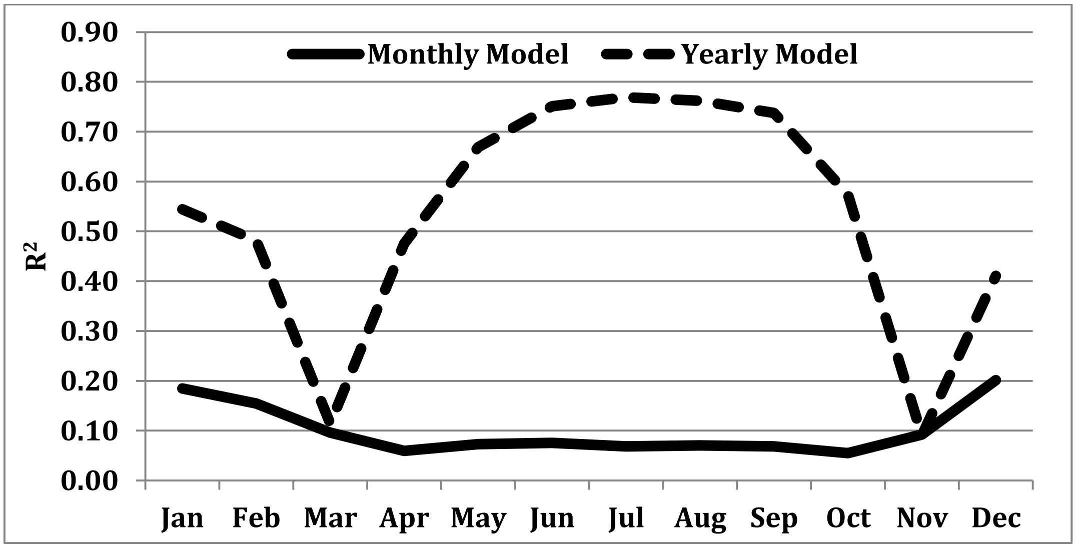

Figure 2. 10. R^2 values of monthly versus yearly model for Natural Gas.

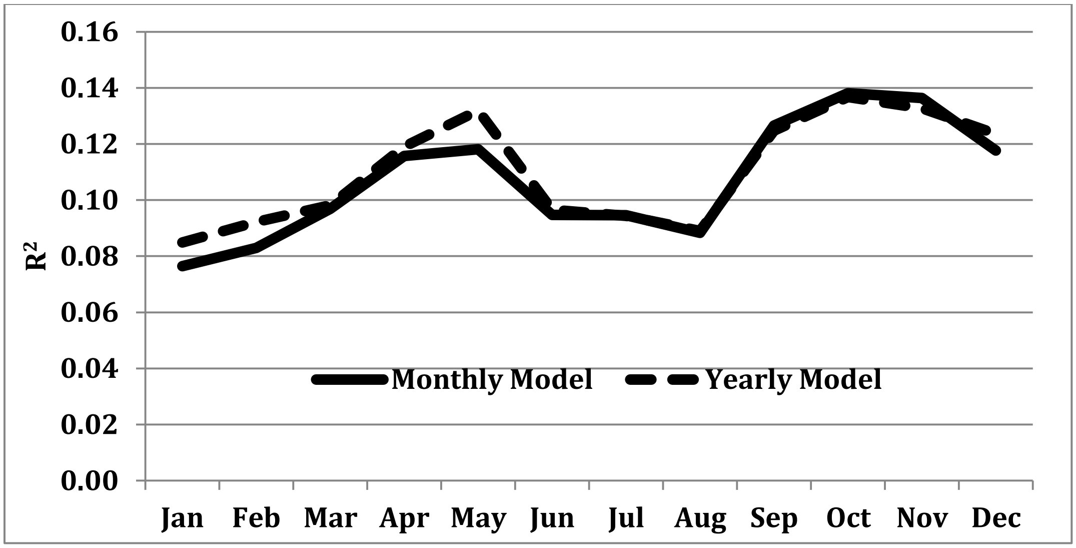

Figure 2. 11. R^2 values of monthly versus yearly model for Water.

2.4. Conclusion

In our work we have shown that publically available data can be used to model residential utility usage in the absence of highly private utility data. We have demonstrated that accounting for billing lag when attributing usage to individual months is of critical importance, particularly for statistical modeling of monthly usage. The accuracy of such models can be diminished if billing lag is not addressed as it can cause significant deviations.

Predicting utility usage by month can allow utilities plan for changes in demand with respect to changes in demographic and structural characteristics in their service territories. This exercise is helpful for planning for new power generation capacity as well as peak demand. Further, predicting monthly utility usage can be translated into energy-use intensity per squarefoot which can be used as a first-cut metric for energy-efficiency planning and targeting.

We acknowledge that our approach has limitations whose extent remains to be explored. The demographic dataset has information on only registered voters and no information at all on children who are younger than 18. We understand that registered voter information may not be full representative of occupants' profile and was used as a proxy. Further, there may incorrect information in the publically available data as collected by property tax assessors and it cannot be known *a priori*. Moreover it is unknown how accurately our models based on the Gainesville sample can predict utility usage for houses outside that are outside our sample, e.g., other houses in Gainesville, Florida or other regions. Collection of further data in other locations can shed light on the extent of our models' accuracy. We understand that model parameters and accurucies may be different if applied to other geographic regions.

Comparing these results to Department of Energy's Residential Energy Consumption Survey (RECS), which harbors aggregate residential energy data for different regions in the US., can generate further insight on our accuracy as well as provide suggestions on how better RECS can be designed and collected. This prospective sample/out-of-sample analysis can prove useful in understanding specific limitations of our models and the independent variables of interest. Future work should also conduct a power analysis to understand the necessary sample size that can produce statistically significant results.

Even though energy-use intensity per squarefoot can be used as a fist-cut metric for EE planning and targeting, it provides limited insight on what type of interventions should be considered for a given house or what kind of utility savings potential a house has. We address this in the following chapters.

The regression models built have significant explanatory power in illustrating the utility usage that can be used by policy makers and third-party developers and operators engaged in energy efficiency and real estate businesses for strategic planning. Aggregate data like RECS does not allow for profiling individual houses and using statistical models based on publically available data can establish the first step to examine the geographic variations in utility usage for large regions. This will pave a path to study what physical and demographic factors drive usage and how they should be treated to promote energy efficiency deployment. Our next study explores this problem to deduce energy-efficiency insight from using usage and publically available information.

2.5. References

1. M. F. Fels, “PRISM: An Introduction” Center for Energy and Environmental Studies, Princeton University, Princeton, NJ 08544. 1986.
2. M. Goldberg , “A Geometrical Approach to Nondifferentiable Regression Models as Related Methods for Assessing Residential Energy Conservation”. Ph.D. Thesis. Department of Statistics, Princeton University, Report No. 142. Center for Energy and Environmental Studies, Princeton, NJ, 1982.
3. A. C. C. MacSleyne, “Residential Energy Consumption and Conservation Programs: A Systematic Approach to Identify Inefficient Houses, Provide Meaningful Feedback, and Prioritize Homes for Conservation Intervention”. Ph.D. Thesis. Department of Engineering and Public Policy, Carnegie Mellon University. 2007
4. National Academy of Sciences, “Real Prospects for Energy Efficiency in the United States”. The National Academies Press, Washington, D.C. 2010
5. American Council for an Energy-Efficient Economy, “State Energy Efficiency Resource Standard (EERS) Activity”. Washington, D.C. 2011
6. Institute for Electric Efficiency, The Edison Foundation, “Summary of Ratepayer-Funded Electric Efficiency Impacts, Budgets, and Expenditures”. 2012
7. A. B. Jaffe, R. N. Stavins, “Energy-Efficiency Investments and Public Policy”. *The Energy Journal*. 1994. Vol. 15. No.2.
8. C. Neme, M. Gottstein, and B. Hamilton, “Residential Efficiency Retrofits: A Roadmap for the Future”. Montpelier, VT: Regulatory Assistance Program. 2011
9. D. Lee, “Better Buildings 2.0”. Presentation at Building America Residential Energy Efficiency Meeting, Denver, CO, July 20, 2010.
10. K. Palmer, M. Walls, M. Gordon and T. Gerarden, “Assessing the Energy-Efficiency Information Gap: Results from a Survey of Home Energy Auditors”. Resources for the Future. Washington, D.C. 2011
11. S. Nowak, M. Kushler, M. Sciortino, D. York, P. Witte, “Energy Efficiency Resource Standards: State and Utility Strategies for Higher Energy Savings”. American Council for an Energy-Efficient Economy. Washington, DC. 2011

12. D. Hynek, “Regression Modeling to Analyze Apartment Space Heating Demand and the Influence of Electrical Use Diversity”. American Council for an Energy-Efficient Economy. Washington, DC. 2006

13. M. C. Fuller, C. Kunkel, M. Zimring, I. Hoffman, K. L. Soroye and C. Goldman, “Driving Demand for Home Energy Improvements: Motivating residential customers to invest in comprehensive upgrades that eliminate energy waste, avoid high bills, and spur the economy”. Lawrence Berkeley National Laboratory. 2010

14. K. Steemers and G. Y. Yun, “House Energy Consumption: A Study of the Role of Occupants”. *Building Research & Information*. 2009, 37 (5-6) pp. 625-637.

15. G. Y. Yun and K. Steemers, “Behavioral, Physical and Socio-economic Factors in House Cooling Energy Consumption”. *Applied Energy*. 2011, 88, pp. 2191-2200.

16. V. Oikonomou, F. Becchis, L. Steg and D. Russolillo, “Energy Saving and Energy Efficiency Concepts for Policy Making”. *Energy Policy*. 2009, 37, pp. 4787-4796.

17. J. Min, Z. Hausfather and Q. F. Lin, “A High-resolution Statistical Model for Residential Energy End Use Characteristics for the United States”. *Journal of Industrial Ecology*. 2010, Vol. 14, Number 5.

18. G. D. Jacobsen, and M. J. Kotchen, “Are Building Code Effective at Saving Energy? Evidence from Residential Billing Data in Florida”. The National Bureau of Economic Research Working Paper No. 16194. 2010

19. M. F. Fels, K. Kissock, M. A. Marean and C. Reynolds, “PRISM (Advanced Version 1.0) Users’ Guide”. Center for Energy and Environmental Studies. Princeton University. Princeton, NJ 08544. 1995.

20. U.S. Department of Energy. http://www.eia.gov/tools/faqs/faq.cfm?id=97&t=3 (Accessed in August 2012).

21. American Water Works Association, 2012. http://www.drinktap.org/consumerdnn/Default.aspx?tabid=85 (Accessed in August 2012)

2.6. APPENDIX

Table A1. Definitions of physical property variables.

Variable	Definition
Land value ($)	The Land Value is the assessed value of the land without an agricultural classification
Building value ($)	The Building Value is the value of the major structures on the property.
Misc. value ($)	The Miscallaneous Value is the value of the miscellaneous improvements on the property.
Property tax amount ($)	The Property Tax Amount is the property tax liability amount due to be paid. It is computed by multiplying the Taxable Value of the property by the Millage Rates of each of the applicable Taxing Authorities.
Property value ($)	Building value plus misc. value
Total property value($)	Property value plus land value
Actual area (sqft)	Actual Area is the number of square feet for any area or subarea of a building structure. Usually the Base or main structure is considered heated and certain types of subareas may be heated in full or in part.
Heated area (sqft)	Heated Area is the number of square feet for all buildings on the property that is considered to be enclosed and subject to heating or cooling.
Age of the building	2012 minus the year that a building on the property was originally constructed. The year 1900 is used when no year of construction is on file.
# of bedrooms	The number of bedrooms of the existing structure on the property.
# of bathrooms	The number of bathrooms of the existing structure on the property.
# of stories	The number of stories of the existing structure on the property. Stories may be defined as full or half story.
Fuel dummy	Binary variable for the fuel used for heating (1= electric heating, 0=natural gas heating)
Natural gas account dummy	Binary variable for whether the house receives natural gas service (1=yes, 0=no)

Table A2. Definitions of demographic characteristic variables

Variable	Definition
Average age of occupants	The average age of the occupants in the building
Average years of occupancy	Sum of years of occupancy divided by the number of occupants. Occupancy is calculated by subtracting the voter registration date from 2012.
# of teenagers	Teenage is the number of occupants who are younger than 20 and older than 17
# of adults	Adult is the number of occupants who are older than 19 and younger than 60
# of seniors	Senior is the number of occupants who are older than 59
# of republicans	Republican is the number of occupants registered as republican
# of democrats	Democrat is the number of occupants registered as democrat
# of males	Male is the number of male occupants
# of females	The number of female occupants

3. Statistical Modeling of Residential Energy-Efficiency Parameters

Abstract

In this study we model and examine the energy efficiency profile of individual single-family houses in our sample (n=7,091). For this we use Princeton Scorekeeping Method (PRISM) which processes historical weather data and monthly utility usage data as inputs using an iterative regression approach to compute three energy efficiency parameters: 1) baseload consumption for end-uses which do not change with weather, e.g., lighting, refrigerator, water heater; 2) heating/cooling slope which is a function of the building shell insulation and the efficiency of the heating/cooling unit; 3) reference temperature which is the outside temperature at which the house turns on heating/cooling. These parameters make up the normalized annual consumption (NAC). We then proceed to regress these parameters against the publically available data to study the extent we can extract statistical insight for residential energy efficiency profiling using publically available information.

3.1. Introduction

This study explores the explanatory power of publically available data on house energy-efficiency parameters as computed by PRISM (Princeton Scorekeeping Method).

We use PRISM to determine the efficiency and consumption profile for each single-family house (n=7,091). PRISM uses daily weather data and monthly utility consumption, (e.g., electricity, natural gas and heating oil) as inputs and estimates baseload/appliance consumption, ambient temperature (thermostat setting), and the thermal integrity/efficiency of the house structure. These parameters constitute a weather-adjusted normalized annual consumption (NAC) as computed by PRISM and is explained in the next section. The next section gives an overview on what underlies PRISM and several case studies.

Processing the datasets used in Chapter 2, statistical models are developed to determine the relationships between the structural and demographic house characteristics and the PRISM output parameters on house baseload consumption, thermostat setting, and thermal integrity for both heating and cooling models.

Disaggregating utility-usage information into three simple parameters as computed by PRISM has distinct advantages in analyzing a utility customer base and its potential energy-efficiency distribution by consumer, i.e., savings potential by energy-efficiency measure across a service territory. These three parameters shed insight on both behavioral and engineering aspects, (e.g., thermostat setting, structural aspects and thermal integrity) of residential buildings. Further, separating heating/cooling-related consumption and baseload usage is insightful in what type of energy-efficiency interventions (e.g., insulation or replacing appliances), can be viable for specific users. Additionally, calculating weather-adjusted normalized annual consumption (NAC) can help establish benchmarks and pinpoint outliers that may be of further interest in energy-efficiency outreach as executed by electric utilities.

Being able to use publically available data and accurately predict structure thermal integrity, baseload consumption and thermostat setting is valuable because every structure has different physical and demographic characteristics and only effective diagnostics of structural and behavioral characteristics of a house can lead to formulation of an intelligent intervention. This is typically executed through a home audit performed by a home-improvement professional. Some houses may need to replace their appliances whereas some may need a building-shell insulation. Some may need a smart thermostat that can regulate the thermostat setting over different times of day. For some houses it may be not an energy-efficiency issue but a comfort issue, i.e., not being able to economically afford to set the indoor temperature at a comfortable level. An intervention in such case is a social welfare issue than a purely economic one. Only disaggregating utility usage into different end-uses can facilitate diagnosing residential energy issues. Using statistical models based on publically available data on individual houses allows this diagnostic exercise to scale up for large regions which provides critical input for utilities and policy-makers to develop analytically-driven energy-efficiency targeting strategies.

It is crucial to underscore the importance of data availability and prioritize collection of certain data if resources are available. Having studied Residential Energy Consumption Survey (RECS), Carlson et al. (2013) emphasize that there can be significant differences between average residential electricity consumption and actual residential consumption. They further assert that making decisions based on RECS average data can be questionable as the data tend to overestimate the number of contributing appliances in a house.

Ndiaye and Gabriel (2011) conducted a principal component analysis (PCA) using 59 predictors to model electricity usage using a sample gathering energy audit data of 62 houses in Ontario, Canada. Their sample included data from phone surveys, home energy audits and smart meter readings. The PCA reduced the number of predictors to 9 which included: 1) the number of occupants in the house; 2) house status (owned vs. rented); 3) average annual number of weeks of vacation taken away from the house by the family; 4) type of fuel used in the pool heater; 5) type of fuel used in the space heating system; 6) type of fuel used in the domestic hot water system; 7) presence or not of an air conditioning system; 8) type of air conditioning system and 9) number of air changes per hour at 50Pa measured via a blower door test. The resulting R^2 was 79%.

Kavousian et al. (2013) use 10-minute interval smart meter data over the course of 238 days in 2010 for 1,628 houses located in the U.S. to determine the demographic and structural house variables to model electricity consumption. The smart meter data were supplemented by a 114-question survey. The major categories of predictors were: 1) weather and location; 2) physical characteristics of the building; 3) appliance and building stock; 4) occupancy and occupants’ behavior towards energy consumption. The models’ R^2 varied between 43% and 68%.

Benchmarking studies are typically encountered in the literature that rely solely on metrics like energy-use intensity (kWh/ft^2) and do not attempt to disaggregate different end-uses. Chun (2011) summarizes the advantages and disadvantages into different techniques encountered in the literature (i.e., simple normalization, ordinary least squares, stochastic frontier analysis and data envelopment analysis).

Kavousian and Rajagopal (2013) propose a stochastic energy-efficiency frontier method (SEEF) that they claim is superior to other benchmarking methods as they treat energy consumption stochastically. SEEF uses an algorithm to determine the functional form of the frontier, identify the probability distribution of efficiency score of each building using measured data, and rank building based on their energy-efficiency (Kavousian and Rajagopal, 2013). They use smart meter data for 307 residential buildings in the U.S, collected between June and September 2010, to illustrate their work.

Brecha et al (2011) used a 1,134-house sample in Yellow Springs, OH that consisted of utility usage information for 2006-2008 and the structural characteristics as given in property tax records. They also conducted “light” house audits in the houses and collected information such as window and wall sizes, R-values for wall, slab/foundation, window and ceiling insulation, and efficiency for HVAC equipment. While such information is highly insightful, it is impractical to collect at a large scale since an audit requires significant time and resources to complete. They broke utility usage into heating/cooling-related consumption and baseload. Analyzing the audit data was useful in estimating potential savings for different efficiency interventions., e.g., sealing leaks or insulating the attic. However, an attempt to establish statistical relationships between such estimates, utility usage information and property tax information was absent. Our study targets exactly these missing pieces in filling the hiatus between public records and energy efficiency profiling to overcome data availability issues.

The motivation behind this study is twofold: 1) only load disaggregation can allow for conducting proper energy-efficiency diagnostics – two houses with the same aggregate utility usage may have different load profiles and different energy-efficiency intervention potential, which can be behavioral or engineering-based, and only separating the end-uses allows for this diagnostics; 2) load disaggregation requires analytical rigor and access to highly private utility usage data; predicting disaggregated loads using only publically available information on structural and demographic house characteristics helps overcome this issue.

3.2. PRISM

PRISM, developed by the Center for Energy and Environmental Studies at Princeton University in 1978, uses daily temperature data from which heating and cooling degree-days are calculated, and monthly utility meter readings for utility consumption as inputs to determine the weather-adjusted index for annual consumption which is called Normalized Annual Consumption (NAC). In essence, NAC is the annual utility consumption for a given year with average weather. PRISM is typically used by researchers and energy-efficiency program managers to compute the effects of energy-efficiency practices across houses and define ways to implement house-retrofit measures more cost effectively (Fels et al., 1995).

PRISM has a variety applications and is widely used for separating utility usage, e.g., electricity, natural gas or heating oil, into disaggregated end-uses: baseload usage, heating/cooling slope and thermostat setting. PRISM has been used specific applications such as tracking retrofit savings (Mills et al., 1987) or scorekeeping for electricity conversation programs (Dutt and Fels, 1989; Gregory, 1987; Hirst, 1986; Rodberg, 1986). Fels and Reynolds (1992) used PRISM for analyzing New York State Energy Research and Development Authority’s (NYSERDA) multifamily conservation program. A similar study was executed by Goldman and Ritschard (1986) for San Francisco Housing Authority to assess energy conservation in public housing. Oak Ridge National Laboratory used PRISM to evaluate national weatherization efforts (Brown et al., 1993). PRISM is not only used for residential buildings or large samples – Haberl et al. (1989) used PRISM to conduct a campus-wide energy performance analysis for Princeton University.

In our study we use monthly natural gas usage data as input for the PRISM heating model to compute $NAC_{\text{natural gas}}$, and monthly electricity usage data for the PRISM cooling model to compute $NAC_{\text{electricity.}}$

NAC is estimated using a three-parameter model that is a function of thermal integrity of the building (Heating/Cooling Slope – HS/CS), appliance-level baseload consumption (BL), and the interior-temperature setting (Reference Temperature – RT). Baseload natural gas consumption is derived from the heating model and is denoted as $BL_{\text{natural gas}}$. Similarly, baseload electricity consumption is derived from the cooling model and is denoted as $BL_{\text{electricity}}$. Heating slope comes from the heating model as denoted as $HS_{\text{natural gas}}$ whereas cooling slope comes from the cooling model and is denoted as $CS_{\text{electricity}}$. $RT_{\text{natural gas}}$ and $RT_{\text{electricity}}$ come from the heating and cooling models, respectively.

Generally, a house's heating/cooling system is operated when the outdoor temperature (T_{out}) goes below/above a certain level (Reference Temperature, RT), and for each incremental degree change in temperature a constant amount of fuel (electricity, fuel oil or natural gas) (the heating/cooling slope (HS/CS) is consumed (Fels, 1986). Hence, the fuel consumed is linearly proportional to (RT – T_{out}) and the constant HS/CS represents the house's effective heat-loss (or gain) rate. Further the house may use a constant amount of fuel per day (the base level BL) independent of T_{out}. This is treated as the baseload of the building and is attributed to appliance-level consumption. Thus, PRISM defines the normalized annual consumption for electricity, $NAC_{\text{electricity}}$, as:

$$NAC_{\text{electricity}} = BL_{\text{electricity}} + CS_{\text{electricity}} \times CDD(RT_{\text{electricity}}) + \varepsilon$$

Where $CDD(RT_{\text{electricity}})$ is the number of cooling degree days for a given $RT_{\text{electricity}}$ in a given year.

Similarly, normalized annual consumption for natural gas, $NAC_{natural\ gas}$, is defined as:

$$NAC_{natural\ gas} = BL_{natural\ gas} + CS_{natural\ gas} \times HDD(RT_{natural\ gas}) + \varepsilon$$

Where $HDD(RT_{natural\ gas})$ is the number of heating degree days for a given $RT_{natural\ gas}$ in a given year.

The derivation of this equation allows the interpretation of the three parameters: The reference temperature RT, which varies from building to building, is likely to be affected by the indoor temperature T_{in}, which is typically set by a thermostat, and intrinsic gains (e.g., heat produced by appliances and occupants, and the sun). The heat loss (and gain) rate $HS_{natural\ gas}/CS_{electricity}$ is governed by conductive and infiltration heat losses/gains as well as the furnace efficiency (Fels, 1986). The base level consumption BL is determined by the amount of fuel consumed by appliances. ε is the random error term that cannot be explained by the regression equation that is solved by ordinary least-squares linear regression technique. Using an iterative approach based on Newton's method (Goldberg, 1982), PRISM solves for the three parameters that best explain changes in fuel consumption.

3.3. Data

The data set used in the previous section (Chapter 2) was also used in this study. However, only single-family houses that use both natural gas and electricity were used (n=7,091). These houses use natural gas as the primary heating source, and electricity for cooling and other appliance-level end-uses. Therefore the heating model uses natural gas usage and the cooling model uses electricity usage. 7,091-house monthly utility usage data for 2009-2011 used in Chapter 2 was processed by PRISM. The output baseload consumption for natural gas can be attributed to end-uses like cooking and possibly clothes drying and water heaters, whereas electricity baseload consists mostly of appliance-level consumption (e.g., lighting, refrigerator, TV).

Because of missing or inconsistent utility entries in the original dataset of 7,091 houses PRISM ran successfully only on 5,243 houses. Inconsistent readings can be due to change in occupants or retrofit activities that were not accounted for in the sample.

The percentile values of the independent variables of the physical and demographic characteristics in the final sample (n=5,243) are given in Tables 1 and 2. The median house had $105K in building value, $30K in land value, $1.8K in miscellaneous value and paid $1.8K of property tax. Further, the median house is single story, has 3 bedrooms, 2 bathrooms, is 33 years old and has an actual area of 2,200 sqft only 1,600 sqft of which is heated.

For the demographic variables, the median values for number of occupants, seniors, adults and teenagers, are 2, 1, 1 and 0, respectively[9]. Additionally, the median of the average age of occupants is 51.8 whereas the median average years of occupancy is 16. Moreover, The median number of males, females, democrats and republicans are 1, 1, 1, and 0, respectively.

Table 3. 1. Percentiles of structural house characteristics

			Percentiles		
	5th	**25th**	**50th**	**75th**	**95th**
Land value ($1000s)	13.0	25.0	30.0	35.0	60.0
Building value ($1000s)	53.6	82.0	105.0	138.6	218.3
Misc. value ($1000s)	0.2	0.9	1.8	3.6	11.6
Tax amount ($1000s)	0.5	1.0	1.8	2.7	4.8
Actual area (1000sqft)	1.2	1.7	2.2	2.7	3.8
Heated area (1000sqft)	1.0	1.3	1.6	2.1	2.9
Age of the building	11	23	33	37	43
# of bedrooms	2	3	3	3	4
# of bathrooms	1	2	2	2	3
# of stories	1	1	1	1	2

[9] Teenage is the number of occupants who are younger than 20 and older than 17. Adult is the number of occupants who are older than 19 and younger than 60. Senior is the number of occupants who are older than 59.

Table 3. 2. Percentiles of demographic house characteristics

	Percentiles				
	5th	25th	50th	75th	95th
Avg. age of occup.	30	41	52	64	82
Avg. years of occup.	4	10	16	25	41
Total # of occupants	1	1	2	2	4
# of teenagers	0	0	0	0	0
# of adults	0	0	1	2	3
# of seniors	0	0	1	1	2
# of Republicans	0	0	0	1	2
# of Democrats	0	0	1	2	3
# of males	0	0	1	1	2
# of females	0	1	1	1	2

Correlations between structural and demographic characteristics were studied primarily to verify quality of the data at hand.

Structural house characteristics that pertain to the size (e.g., square footage and number of rooms) and value (e.g., building value, property tax amount) of the building are positively correlated between each other (Table 3). Age of the building is negatively correlated with these characteristics.

Adults and seniors tend not to live together (ρ=-0.56). The occupants in the sample are mostly adults and seniors because we have no data on children as extracted from voter registration records. Ostensibly, seniors are than older adults, thus the number of adults in a house drives the average occupant age down (ρ= -0.70) whereas the number of seniors increases it (ρ=0.68) (Table 3). Further, number of seniors is positively correlated with higher number of years of occupancy (ρ=0.44).

Democrats and Republicans tend to live separately (ρ=-0.47).

Table 3. 3. Correlations between structural and demographic house characteristics

	Land value	Building value	Misc. value	Tax amount	Actual area	Heated area	Age of the building	# of bedrooms	# of bathrooms	# of stories	Avg. age of occupants	Avg. years of occupancy	Total # of occupants	# of teenagers	# of adults	# of seniors	# of Republicans	# of Democrats	# of males	# of females
Land value ($1000s)	1.00																			
Bldg value ($1000s)	0.57	1.00																		
Misc. value ($1000s)	0.38	0.56	1.00																	
Tax amount ($1000s)	0.56	0.86	0.53	1.00																
Actual area (1000sqft)	0.54	0.91	0.58	0.78	1.00															
Heated area (1000sqft)	0.54	0.89	0.59	0.76	0.95	1.00														
Age of the building	0.00	-0.27	0.00	-0.33	-0.13	-0.08	1.00													
# of bedrooms	0.28	0.53	0.36	0.43	0.60	0.62	-0.03	1.00												
# of bathrooms	0.40	0.70	0.46	0.61	0.71	0.71	-0.19	0.51	1.00											
# of stories	0.21	0.31	0.18	0.27	0.33	0.35	0.05	0.22	0.28	1.00										
Avg. age of occupants	0.03	0.07	0.06	-0.11	0.11	0.11	0.10	0.05	0.05	0.01	1.00									
Avg. years of occupancy	0.04	0.06	0.06	-0.13	0.12	0.12	0.21	0.05	0.05	0.05	0.62	1.00								
Total # of occupants	0.03	0.12	0.10	0.07	0.14	0.16	0.03	0.17	0.09	0.04	-0.28	-0.17	1.00							
# of teenagers	0.01	0.02	0.01	0.02	0.02	0.02	0.00	0.04	0.00	0.01	-0.14	-0.10	0.17	1.00						
# of adults	-0.02	0.00	0.02	0.09	0.00	0.00	-0.05	0.05	0.01	0.01	-0.70	-0.44	0.72	0.08	1.00					
# of seniors	0.06	0.14	0.10	-0.04	0.18	0.19	0.11	0.13	0.10	0.03	0.68	0.44	0.17	-0.07	-0.56	1.00				
# of Republicans	0.07	0.12	0.10	0.11	0.13	0.13	-0.05	0.11	0.13	0.04	-0.09	-0.08	0.29	0.05	0.23	0.02	1.00			
# of Democrats	-0.03	0.00	0.02	-0.05	0.03	0.04	0.11	0.07	-0.02	0.01	-0.04	0.07	0.52	0.06	0.29	0.20	-0.47	1.00		
# of males	0.08	0.13	0.10	0.11	0.15	0.16	0.04	0.14	0.11	0.07	-0.24	-0.14	0.69	0.10	0.51	0.10	0.24	0.29	1.00	
# of females	-0.02	0.04	0.04	-0.01	0.05	0.07	0.00	0.10	0.03	0.00	-0.12	-0.06	0.65	0.13	0.44	0.15	0.16	0.41	-0.05	1.00

3.4. PRISM Simulations

As mentioned in the previous section, PRISM ran successfully on 5,243 houses of 7,091, and only 3,440 of those houses had good fits (R^2>0.70) for both heating and cooling models. Although, PRISM suggests using R^2>0.70 as a good-fit benchmark, all simulations (n=5,243) were kept for further examination. Monthly natural gas and electricity usage values were used for heating and cooling models, respectively. The percentile values of PRISM output parameters (i.e., reference temperature (RT), baseload consumption (BL) and heating/cooling slope (HS/CS) (cooling slope for electricity, heating slope for natural gas usage), normalized annual consumption (NAC), and the associated standard errors) are given in Tables 4 and 5 (n=5,243).

Table 3. 4. Percentiles of PRISM parameters.

	Percentiles				
	5th	**25th**	**50th**	**75th**	**95th**
$R^2_{natural\ gas}$	0.46	0.82	0.88	0.91	0.93
$R^2_{electricity}$	0.23	0.65	0.82	0.90	0.95
$RT_{natural\ gas}$(°C)	11.00	13.00	17.56	19.00	23.63
$RT_{electricity}$(°C)	15.60	19.00	20.92	22.89	25.63
$BL_{natural\ gas}$ (therm/day)	0.12	0.23	0.34	0.47	0.71
$BL_{electricity}$ (kWh/day)	7.55	12.62	17.61	24.70	40.69
$HS_{natural\ gas}$ (therm/degree-day)	0.03	0.21	0.33	0.51	0.93
$CS_{electricity}$ (kWh/degree-day)	1.42	2.50	3.46	4.74	7.80
$NAC_{natural\ gas}$ (therm/year)	115	206	284	379	567
$NAC_{electricity}$ (kWh/year)	4,399	7,380	10,040	13,336	20,176

Table 3. 5.Percentiles of PRISM parameters' variances

	Percentiles				
	5th	**25th**	**50th**	**75th**	**95th**
$RT_{natural\ gas}$ Var	0.79	1.37	2.04	3.31	21.23
$RT_{electricity}$ Var	0.50	1.17	2.43	5.81	28.09
$BL_{natural\ gas}$ Var	0.001	0.002	0.004	0.009	0.028
$BL_{electricity}$ Var	0.35	1.16	2.60	6.06	22.71
$HS_{natural\ gas}$ Var	0.0001	0.0015	0.0046	0.0139	0.0640
$CS_{electricity}$ Var	0.07	0.22	0.54	1.57	13.38
$NAC_{natural\ gas}$ Var	13,119	42,319	80,906	143,603	321,881
$NAC_{electricity}$ Var	17,264	49,364	100,724	222,458	693,354

The median $RT_{electricity}$ is computed as 20.9°C and $RT_{natural\ gas}$ as 17.6°C for heating. The median value for $BL_{electricity}$ is 17.6kWh/day and 0.34therm/day for $BL_{natural\ gas}$. Further, median $CS_{electricity}$ and $HS_{natural\ gas}$ are 3.46kWh/degree-day and 0.33/therms/degree-day, respectively. The median $NAC_{natural\ gas}$ is 284therms/year and $NAC_{electricity}$ is 10,040Wh/year (Table 4).

The percentiles of PRISM parameters were investigated to understand how disaggregated end-uses vary from house to house. Additionally, the next section (Chapter 4) uses median PRISM values for a benchmarking exercise to examine savings potential by end-use and house.

Table 6 shows the correlations between the PRISM output parameters and the associated variances. A negative correlation between $HS_{natural\ gas}$ and $RT_{natural\ gas}$ (ρ=-0.48) suggest that houses with less thermal integrity (high $HS_{natural\ gas}$) tend to decrease their thermostat settings for heating to conserve energy.

Output parameters that directly constitute NAC (i.e., BL, RT and CS/HS) are positively correlated with NAC for both natural gas and electricity. Also, the output parameters are positively correlated with their individual variances (Table 6) (i.e., the larger the parameter the larger the associated uncertainty). This verifies that PRISM produced plausible output parameters.

Table 3. 6. Correlations between PRISM output parameters and the associated variances

	RT$_{\text{natural gas}}$	RT$_{\text{electricity}}$	BL$_{\text{natural gas}}$	BL$_{\text{electricity}}$	HS$_{\text{natural gas}}$	CS$_{\text{electricity}}$	NAC$_{\text{natural gas}}$	NAC$_{\text{electricity}}$	RT$_{\text{natural gas}}$ **Var**	RT$_{\text{electricity}}$ **Var**	BL$_{\text{natural gas}}$ **Var**	BL$_{\text{electricity}}$ **Var**	HS$_{\text{natural gas}}$ **Var**	CS$_{\text{electricity}}$ **Var**	NAC$_{\text{natural gas}}$ **Var**	NAC$_{\text{electricity}}$ **Var**
RT$_{\text{natural gas}}$	1.00															
RT$_{\text{electricity}}$	0.14	1.00														
BL$_{\text{natural gas}}$	-0.09	-0.18	1.00													
BL$_{\text{electricity}}$	0.05	0.03	0.28	1.00												
HS$_{\text{natural gas}}$	-0.48	0.00	0.18	0.18	1.00											
CS$_{\text{electricity}}$	0.04	0.21	0.01	0.09	0.01	1.00										
NAC$_{\text{natural gas}}$	0.15	-0.04	0.58	0.27	0.31	0.02	1.00									
NAC$_{\text{electricity}}$	-0.05	-0.23	0.37	0.91	0.24	0.04	0.35	1.00								
RT$_{\text{natural gas}}$ **Var**	0.21	0.04	0.00	0.02	-0.06	0.04	-0.05	-0.01	1.00							
RT$_{\text{electricity}}$ **Var**	0.04	-0.23	-0.02	-0.01	-0.02	-0.01	-0.02	-0.04	0.00	1.00						
BL$_{\text{natural gas}}$ **Var**	0.05	0.01	0.18	0.04	0.00	0.00	0.17	0.03	0.13	0.00	1.00					
BL$_{\text{electricity}}$ **Var**	0.02	-0.13	0.00	0.06	0.01	-0.01	0.03	0.02	0.00	0.45	0.00	1.00				
HS$_{\text{natural gas}}$ **Var**	-0.15	0.01	0.12	0.11	0.72	0.01	0.07	0.09	-0.01	0.00	0.00	0.00	1.00			
CS$_{\text{electricity}}$ **Var**	0.02	0.13	-0.01	0.08	0.00	0.66	-0.02	0.03	0.01	0.00	0.00	0.00	0.00	1.00		
NAC$_{\text{natural gas}}$ **Var**	0.15	-0.01	0.52	0.26	0.27	0.03	0.89	0.32	0.02	-0.01	0.44	0.02	0.08	-0.01	1.00	
NAC$_{\text{electricity}}$ **Var**	0.05	0.03	0.13	0.57	0.25	0.12	0.13	0.54	0.02	0.06	0.00	0.15	0.28	0.11	0.15	1.00

Table 7 shows the correlations between PRISM parameters and the structural house characteristics. House size is positively correlated with $BL_{natural\ gas}$ and $BL_{elecricity}$. This may be because larger houses tend to have more appliances. House size is also positively correlated with $HS_{natural\ gas}$ and $CS_{electricity}$, which are a function of the surface area of the building. Further, larger houses also are positively correlated with larger $NAC_{electricity}$ and $NAC_{natural\ gas}$ (Table 7). These correlations confirm that PRISM generated plausible output parameters.

No strong correlations were found between PRISM parameters and demographic house characteristics (Table 8).

Table 3. 7. Correlations between PRISM output parameters and structural house characteristics.

	RT$_{\text{natural gas}}$	**RT**$_{\text{electricity}}$	**BL**$_{\text{natural gas}}$	**BL**$_{\text{electricity}}$	**HS**$_{\text{natural gas}}$	**CS**$_{\text{electricity}}$	**NAC**$_{\text{natural gas}}$	**NAC**$_{\text{electricity}}$	**Land value**	**Building value**	**Misc. value**	**Tax amount**	**Actual area**	**Heated area**	**Age of the building**	**# of bedrooms**	**# of bathrooms**	**# of stories**
RT$_{\text{natural gas}}$	1.00																	
RT$_{\text{electricity}}$	0.14	1.00																
BL$_{\text{natural gas}}$	-0.09	-0.18	1.00															
BL$_{\text{electricity}}$	0.05	0.03	0.28	1.00														
HS$_{\text{natural gas}}$	-0.48	0.00	0.18	0.18	1.00													
CS$_{\text{electricity}}$	0.04	0.21	0.01	0.09	0.01	1.00												
NAC$_{\text{natural gas}}$	0.15	-0.04	0.58	0.27	0.31	0.02	1.00											
NAC$_{\text{electricity}}$	-0.05	-0.23	0.37	0.91	0.24	0.04	0.35	1.00										
Land value ($1000s)	-0.04	0.02	0.05	0.29	0.22	0.03	0.19	0.30	1.00									
Building value ($1000s)	-0.10	0.00	0.19	0.48	0.31	0.03	0.33	0.50	0.57	1.00								
Misc. value ($1000s)	-0.07	-0.02	0.16	0.52	0.22	0.04	0.23	0.51	0.38	0.56	1.00							
Tax amount ($1000s)	-0.08	-0.02	0.19	0.43	0.24	0.05	0.24	0.45	0.56	0.86	0.53	1.00						
Actual area (1000sqft)	-0.11	0.04	0.19	0.53	0.34	0.05	0.36	0.55	0.54	0.91	0.58	0.78	1.00					
Heated area (1000sqft)	-0.11	0.04	0.19	0.55	0.35	0.05	0.38	0.57	0.54	0.89	0.59	0.76	0.95	1.00				
Age of the building	0.03	0.19	-0.05	0.06	0.12	0.05	0.14	0.02	0.00	-0.27	0.00	-0.33	-0.13	-0.08	1.00			
# of bedrooms	-0.08	0.00	0.16	0.35	0.21	0.01	0.25	0.37	0.28	0.53	0.36	0.43	0.60	0.62	-0.03	1.00		
# of bathrooms	-0.08	-0.02	0.15	0.41	0.23	0.03	0.23	0.43	0.40	0.70	0.46	0.61	0.71	0.71	-0.19	0.51	1.00	
# of stories	-0.03	0.04	0.06	0.21	0.11	0.03	0.11	0.20	0.21	0.31	0.18	0.27	0.33	0.35	0.05	0.22	0.28	1.00

Table 3. 8. Correlations between PRISM output parameters and demographic house characteristics.

	$RT_{natural\ gas}$	$RT_{electricity}$	$BL_{natural\ gas}$	$BL_{electricity}$	$HS_{natural\ gas}$	$CS_{electricity}$	$NAC_{natural\ gas}$	$NAC_{electricity}$	Avg. age of occupants	Avg. years of occupancy	Total # of occupants	# of teenagers	# of adults	# of seniors	# of Republicans	# of Democrats	# of males	# of females
$RT_{natural\ gas}$	1.00																	
$RT_{electricity}$	0.14	1.00																
$BL_{natural\ gas}$	-0.09	-0.18	1.00															
$BL_{electricity}$	0.05	0.03	0.28	1.00														
$HS_{natural\ gas}$	-0.48	0.00	0.18	0.18	1.00													
$CS_{electricity}$	0.04	0.21	0.01	0.09	0.01	1.00												
$NAC_{natural\ gas}$	0.15	-0.04	0.58	0.27	0.31	0.02	1.00											
$NAC_{electricity}$	-0.05	-0.23	0.37	0.91	0.24	0.04	0.35	1.00										
Avg. age of occup.	0.07	0.13	-0.17	-0.07	0.05	0.01	0.13	-0.11	1.00									
Avg. years of occup.	0.03	0.10	-0.09	0.01	0.07	-0.01	0.13	-0.02	0.62	1.00								
Total # of occup.	-0.04	-0.04	0.23	0.28	0.06	0.00	0.14	0.30	-0.28	-0.17	1.00							
# of teenagers	-0.03	-0.02	0.07	0.06	0.02	0.00	0.03	0.08	-0.14	-0.10	0.17	1.00						
# of adults	-0.04	-0.10	0.23	0.18	0.00	-0.01	0.00	0.21	-0.70	-0.44	0.72	0.08	1.00					
# of seniors	0.02	0.09	-0.06	0.08	0.08	0.01	0.16	0.05	0.68	0.44	0.17	-0.07	-0.56	1.00				
# of Republicans	-0.04	-0.10	0.08	0.17	0.02	0.00	0.01	0.21	-0.09	-0.08	0.29	0.05	0.23	0.02	1.00			
# of Democrats	-0.01	0.07	0.12	0.10	0.06	0.00	0.14	0.08	-0.04	0.07	0.52	0.06	0.29	0.20	-0.47	1.00		
# of males	-0.01	-0.04	0.15	0.24	0.05	-0.02	0.09	0.25	-0.24	-0.14	0.69	0.10	0.51	0.10	0.24	0.29	1.00	
# of females	-0.04	-0.02	0.16	0.15	0.04	0.02	0.10	0.16	-0.12	-0.06	0.65	0.13	0.44	0.15	0.16	0.41	-0.05	1.00

3.5. Models

Weighted least squares regression was used to predict the PRISM output parameters using publically available structural and demographic house characteristics as explanatory variables[10]. Each observation represents a PRISM simulation where each individual PRISM simulation has four output variables, i.e., BL, CS/HS, RT and NAC, their associated variances.

PRISM itself uses a least squares regression to fit the best function using utility and weather information. Given the PRISM results inherently contain uncertainty themselves a weighted least squares regression, where the weight allocated to each observation is inversely proportional to their variance, was chosen as a more appropriate regression model. Weighted least squares method is commonly used to account for heteroscedasticity where sub-populations have different variabilites. Variabilities can be measured in variance or any other measure for statistical dispersion. Contrary to traditional least squares method, weighted least squares fits a function where more uncertain observations gets a smaller weight and the regression model minimizes the sum of weighted residuals instead of sum of residuals. Weighted least squares is used only when the analyst can estimate the variability, e.g., variance, for the observations at hand. If the variabilities cannot be estimated the analyst ought to revert to other regression methods.

To clarify, we regressed PRISM output parameters using publically available information on the structural and demographic characteristics households to be able to predict them in the absence of highly private utility usage information that does not always exist in a digital format.

[10] Stata created by StataCorp was used to run the regression models. Originally, stepwise regressions were run. PRISM computes standard errors associated with the output variables. Hence, we ran weighted least squares method using the standard errors which greatly improved the fit of our model.

Table 9 gives an overview for the eight regression models for $RT_{natural\ gas}$, $RT_{electricity}$, $BL_{natural\ gas}$, $BL_{electricity}$, $HS_{natural\ gas}$, $CS_{electricity}$, $NAC_{natural\ gas}$ and $NAC_{electricity}$, and the resulting statistically significant independent variables (p=0.05).

The resulting R^2 values vary between 24% and 69%, $RT_{natural\ gas}$ model having the lowest (24%) and $NAC_{natural\ gas}$ having the highest (69%).

To clarify, PRISM computes the simulations trying to minimize the error in the output NAC value and uses an iterative regressive approach where utility usage and historical weather information are input variables. We used publically available information to predict PRISM output parameters to overcome the necessity of: 1) Gathering highly private utility usage information; 2) Processing the information via PRISM.

Table 3. 9. Beta coefficients for PRISM regressions. All variables are statistically significant at p=0.05.

Independent variables				**Models**							
Min	**Median**	**Max**		$RT_{natural\ gas}$	$RT_{electricity}$	$BL_{natural\ gas}$	$BL_{electricity}$	$HS_{natural\ gas}$	$CS_{electricity}$	$NAC_{natural\ gas}$	$NAC_{electricity}$
0	30	480	**Land value ($1000s)**	-0.03	0.03	-0.002	-0.01	-0.0001		0.39	-12
17	105	726	**Building value ($1000s)**	-0.02	-0.07	0.002	-0.01		-0.02	-0.34	
0	2	31	**Misc. value ($1000s)**	-0.13	0.25		0.68		-0.10	-0.74	213
0	2	17	**Tax amount ($1000s)**	0.37	1.27	0.042	0.48	0.0003	-0.06	4.00	218
1	2	13	**Actual area (1000sqft)**		0.77	-0.095	-1.07	0.0074	1.30	19.25	
1	2	8	**Heated area (1000sqft)**	1.68	4.05	-0.070	3.48	-0.0133	1.75	-6.16	2,091
4	33	72	**Age of the building**	-0.003	0.001	0.00002		0.00002		0.03	
1	3	5	**# of bedrooms**	-0.36	-0.84	0.019	0.50	0.0025	-0.48	3.87	-168
1	2	7.5	**# of bathrooms**	-1.44	-4.02	-0.065	0.81	0.0018		-20.11	635
0	1	3	**# of stories**	0.79	0.51		0.88	-0.0003	-0.33	2.69	
18	52	101	**Avg. age of occupants**		0.04	0.002		-0.0001	-0.02	1.19	-16
0	16	54	**Avg. years of occupancy**	0.11	-0.09	0.001	-0.02	0.0001	-0.01	0.18	
1	2	8	**Total # of occupants**	1.71	-2.75	0.227	1.79	0.0072		134.67	952
0	1	4	**# of seniors**	1.99	-1.75	0.223	2.63	-0.0087	-0.52	121.05	1,506
0	0	2	**# of teenagers**	1.19	0.77			0.0006	0.41	-14.02	273
0	0	6	**# of Republicans**	-1.22	-0.74	-0.011	2.04	0.0022	0.57	-10.71	1,301
0	1	8	**# of Democrats**	-1.25	-0.45	0.026	1.29	0.0016	0.15	-9.97	706
0	1	6	**# of males**		1.11	-0.185		-0.0066		-125.26	
0	1	7	**# of females**		3.13	-0.220	-0.95	-0.0050		-126.35	
			Intercept	10.62	30.55	0.065	-1.63		0.73	-27.99	
			R^2	0.24	0.67	0.54	0.54	0.38	0.65	0.69	0.52
			n	5,243	5,243	5,243	5,243	5,243	5,243	5,243	5,243

3.6. Discussion

Load disaggregation is necessary for energy-efficiency profiling and conducting proper diagnostics to understand what constitutes an effective energy-efficiency intervention for given house. Two houses with the same aggregate use may have different intervention potential on different end-uses and only load disaggregation enables this. Our work shows that conducting a PRISM-driven load disaggregation on monthly utility usage information is computationally light and straightforward.

The generated insight can be helpful in energy-efficiency profiling for large utility customer bases. We found strong correlations between structural, demographic and energy-efficiency characteristics. Our weighted least squares regression results indicate that publically available information on structural and demographic house characteristics can explain the variability in energy-efficiency parameters to varying degrees. While other studies have tried to model residential utility usage incorporating similar variables, our work takes it a step further and explored such variables' power to statistically the energy-efficiency profile of residential buildings. Given the relatively significant R^2 values of our weighted least squares regression models, publically available information can be helpful in assessing residential building stocks' efficiency profile in the absence of private utility data. With our models utility can plan their EE programs with respect to structural and demographic changes in their services territories. Further, utilities can conduct large-scale virtual and non-invasive home audits to assess the energy-efficiency profile of their service territories.

This exercise will further be enriched via quantifying the energy efficiency potential amongst houses in our sample. The PRISM computed efficiency parameters allow estimating the tradeoffs between behavioral and structural energy efficiency interventions. We address these issues in Chapter 4.

3.7. Conclusion

Processing utility usage information using software like PRISM can help building baselines and define efficiency profiles for individual houses in utility service territories. However, utility usage information is highly private and typically not shared with 3rd parties. The motivation of this paper is to explore to which extent publically available information on physical and demographic house characteristics can help understand efficiency profiles of houses in the absence of utility data, which is often the case. This work shows that such publically available information can help predict efficiency parameters, as computed by PRISM, i.e. thermostat setting, thermal integrity, baseload consumption, within varying degrees of explanatory power. Our findings further underscore that availability of data and analytical use thereof are critical for understanding the US building stock for energy efficiency targeting and accelerate energy efficiency deployment. Expanding available datasets to different frontiers such as smart meters, smart thermostats or house audits can strengthen the potential analytical insight we can derive.

Being able to use publically available data and accurately predict structure thermal integrity, baseload consumption and thermostat setting are valuable because every structure has different physical and demographic characteristics and only knowing what constitutes an energy-efficiency problem can lead to formulation of an intelligent intervention. Only disaggregating utility usage into different end-uses can facilitate diagnosing residential energy issues. Using statistical models based on publically available data on individual houses allows this diagnostic exercise to scale up for large regions which provides critical input for utilities and policy-makers to develop analytically-driven energy-efficiency targeting strategies. This in turn will enable decision-makers to assess residential building stocks from an economic as well as a social welfare perspective, and design, implement and evaluate their energy-efficiency programs systematically.

There are limitations to our work. Publically available data may contain incorrect entries that can distort the accuracy of our models. PRISM does not run simulations with incorrect or missing utility usage values. Therefore, we had to remove some of our observations that PRISM could not successfully process from our sample. Moreover, PRISM uses an iterative regression method to calculate the efficiency parameters and their associated variance. Thus, the dependent variables, i.e., the efficiency parameters, are uncertain and we addressed this using a weighted least squares regression. Calculating the efficiency parameters alone do not shed sufficient insight on what kind of savings potential a house has for a given intervention. This exercise ought to be extended using benchmarking. In other words houses need to be compared to a baseline to estimate savings potential. We address this problem in the next chapter.

3.8. References

1. R. J. Brecha, A. Mitchell, K. Hallinan and K. Kissock, "Prioritizing Investment in Residential Energy Efficiency and Renewable Energy – A Case Study for the U.S. Midwest". *Energy Policy*. 2011, Vol. 39, pp. 2982-2992.
2. M. Brown, L. Berry and L. Kinney, "Weatherization Works: An Interim Report of the National Weatherization Evaluation". Report No. ORNL/CON-373, Oak Ridge National Laboratory, Oak Ridge, TN. 1993
3. D. R. Carlson, H. S. Matthews and M. Berges, "One Size Does Not Fit All: Averaged Data on Household Electricity is Inadequate for Residential Energy Policy and Decisions". *Energy and Buildings*. 2013, Vol. 64, pp. 132-144.
4. W. Chun, "Review of Building Energy-use Performance Benchmarking Methodologies". *Applied Energy*. 2011, Vol. 88, pp. 1470-1479.
5. G. Dutt and M. Fels, "Keeping Score in Electricity Conservation Programs. Electricity: Efficient End-Use and New Generation Technologies, and Their Planning Implications". Vattenfall Electricity Congress. Lund University Press, Lund, Sweden. 1989, pp. 353- 388
6. M. F. Fels, "PRISM: An Introduction". Center for Energy and Environmental Studies, Princeton University, Princeton, NJ 08544. 1986
7. M. F. Fels and C. Reynolds, "Energy Analysis in New York City Multifamily Building: Making Good Use of Available Data". Report NYSERDA 93-3, New York State Energy Research and Development Authority, Albany, NY. 1992
8. M. F. Fels, K. Kissock, M. A. Marean and C. Reynolds, "PRISM (Advanced Version 1.0) Users' Guide". Center for Energy and Environmental Studies. Princeton University. Princeton, NJ 08544. 1995
9. M. Goldberg, "A Geometrical Approach to Nondifferentiable Regression Models as Related Methods for Assessing Residential Energy Conservation". Ph.D. Thesis. Department of Statistics, Princeton University, Report No. 142. Center for Energy and Environmental Studies, Princeton, NJ, 1982.

10. C. Goldman and R. Ritschard, "Energy Conservation in Public Housing: a Case Study of the San Francisco Housing Authority". *Energy and Buildings*. 1986, Vol. 9, pp. 89-98.
11. J. Gregory, "Ohio Home Weatherization Assistance Program Final Report". Office of Weatherization, Ohio Dept. of Development, Columbus, OH. 1987
12. J. S. Haberl, S. Englander, C. Reynolds, M. McKay and T. Nyquist, "Whole-Campus Performance Analysis Methods: Early Results from Studies at the Princeton Campus." Proceedings of 6th Annual Symposium on Improving Energy Efficiency in Hot and Humid Climates. Texas A & M University, Dallas, TX. 1989
13. E. Hirst, "Electricity Savings One, Two and Three Years After Participation in the BPA Residential Weatherization Pilot Program". *Energy and Buildings*. 1986, Vol. 9, pp. 45-53.
14. A. Kavousian and R. Rajagopal, "Data-Driven Benchmarking of Building Energy Efficiency Utilizing Statistical Frontier Models*". Journal of Computing in Civil Engineering*. 2013. 10.1061/(ASCE)CP.1943-5487.0000327
15. A. Kavouisian, R. Rajagopal and M. Fischer, "Determinants of Residential Electricity Consumption: Using Smart Meter Data to Examine the Effect of Climate, Building Characteristics, Appliance Stock, and Occupants' Behavior". *Energy*. In Press (2013).
16. A. C. C. MacSleyne, "Residential Energy Consumption and Conservation Programs: A Systematic Approach to Identify Inefficient Houses, Provide Meaningful Feedback, and Prioritize Homes for Conservation Intervention". Ph.D. Thesis. Department of Engineering and Public Policy, Carnegie Mellon University. 2007
17. E. Mills, M. Fels and C. Reynolds, "PRISM: A Tool for Tracking Retrofit Savings". *Energy Auditor and Retrofitter*. 1986, Nov/Dec. pp. 27-34.
18. D. Ndiaye and K. Gabriel, "Principal Component Analysis of the Electricity Consumption in Residential Dwellings". *Energy and Buildings*. 2011, Vol. 43, pp. 446-453.
19. L. Rodberg, "Energy Conservation in Low-Income Homes in New York City: The Effectiveness of House Doctoring". *Energy and Buildings*. 1986, Vol. 9, pp. 55-64.

4. Statistical Modeling of Potential Changes in Utility Usage Due to Energy-Efficiency Interventions

Abstract

This study uses a load-disaggregation model PRISM and its energy efficiency output parameters to determine the savings potential in individual single-family houses (n=5,243) in Gainesvile, FL, studied in the previous chapters, using different interventions, e.g., heating unit replacement, changing the thermostat setting. Every house is compared to the sample's median on baseload heating slope and reference temperature, and the monetary value is computed through calculating the change in annual consumption if the house were to change/improve its efficiency parameter to the sample's median through a hypothetical intervention. This is to research the notion of an energy efficiency reservoir and assess the energy efficiency potential distribution by house and measure. In addition, we quantify the potential rebound amount through profiling houses that set their thermostat higher than median for summer and lower for winter, which also have a positive efficiency potential through changing their heating slope to the sample's median. Finally, we regress the efficiency potential from different interventions against the publically available data to create an algorithm to identify houses with large savings potential for specific interventions.

4.1. Introduction

Energy efficiency (EE) and energy conservation today are recognized as the low-hanging fruit of energy sources (NAS, 2010). In recent years, several states have recognized the potential for energy efficiency to reduce energy consumption and pollutants' emissions, as well as possibly avoiding some new generation construction. Thus, in order to promote energy efficiency, 24 states to-date have enacted Energy Efficiency Resource Standards (EERS) and set reduction targets for energy consumption (ACEEE, 2011). These targets have goals that range between 0% and 2.2% of annual reductions from the baseline (ACEEE, 2011).

To achieve energy efficiency goals, several strategies can be pursued. One of these is the use of demand-side management (DSM) programs. Almost $7 billion was spent in rate-payer-funded DSM programs at a national level in 2011and it is anticipated that a total of $12 billion will be spent by 2020 (IEE, 2012). Although a large number of states are meeting their demand reduction targets, with the coming more aggressive reduction goals, the exercise at hand will become more difficult. The potential benefits of energy efficiency are often unrealized due to market failures and market barriers. Some of these include information barriers, split incentives, hidden costs, transaction costs, high discount rates and heterogeneity among potential adopters (Jaffe and Stavins, 1994). Additionally, unpriced costs and benefits; misconstrued fiscal and regulatory policies; and insufficient and inaccurate information are recognized as market failures (NAS, 2010) Low priority of energy issues; incomplete markets for energy efficiency; limited access to capital (e.g. loans), further constitute market barriers for efficiency deployment (NAS, 2010).

Electric utilities, subject to aforementioned regulations commonly use energy use intensity (kWh/ft^2) as key metric in implementing their energy efficiency programs. Benchmarking studies are typically encountered in the literature, which rely solely on metrics like energy use intensity and do not attempt to disaggregate different end uses. Chun (2011) summarizes the advantages and disadvantages of different techniques encountered in the literature (i.e., simple normalization; ordinary least squares; stochastic frontier analysis and data envelopment analysis). Kavousian and Rajagopal (2013) propose a stochastic energy-efficiency frontier method, which they claim is superior to other benchmarking methods as they treat energy consumption stochastically. They use smart meter data for 307 residential buildings to illustrate their work.

Attempts to model energy use using a combination of utility billing data, smart meter data, publicly available records and home energy audit data are fairly common[11].

[11] See Chapter 3 for an overview of the existing literature.

However, using energy intensity and benchmarking as the primary technique in identifying houses as prime candidates for specific energy-efficiency interventions lacks the necessary level of complexity, as energy efficiency at large is a broader concept than just energy intensity.

Our study processes monthly utility billing data and historical weather data for load disaggregation and determines the baseload consumption, heating/cooling slope and the reference temperature at which the thermostat is set. Extracting these specific energy-efficiency parameters is valuable for locating prime candidates for interventions (e.g., insulation and HVAC unit replacement). Comparing these parameters to a given sample's median values simplifies identifying outliers and targeting them for an energy-efficiency improvement. We then build statistical models to based on publically available information on structural and demographic characteristics of houses to predict the savings potential in dollars per year terms for each house and energy-efficiency parameter, i.e., baseload consumption, heating/cooling slope and reference temperature. Since our models are based on publically available information, this approach overcomes the necessity of access to highly private utility usage data. These models can be used by utilities and 3rd party EE program implementers to provide targeted messaging customized EE feedback to individual houses in a given service territory.

Ehrhardt-Martinez et al. (2010) studied 36 EE feedback programs between 1995 and 2010. These feedback programs include indirect feedback (provided after consumption occurs) and direct feedback (provided real-time). Indirect feedback programs that encompass enhanced billing information, web-based energy audits generated estimated annual percent savings between 3.8% and 8.4%. Direct feedback programs in their study were evaluated to generate 9.2-12.0% savings and used real-time feedback some of which included appliance-level information. These results underscore the importance of feedback programs and how effective they can be. Our statistical models based on publically available information could facilitate a more scalable and analytical EE program design and implementation that are not subject to lack of highly private utility usage data.

We use PRISM software to process the utility billing data and historical weather data to generate disaggregate loads into baseload consumption, heating/cooling slope and reference temperature

PRISM is typically used by researchers and energy-efficiency program managers to compute the effects of energy-efficiency practices across houses and define ways to implement house-retrofit measures more cost effectively (Fels et al., 1995).

PRISM has a variety applications and is widely used for separating utility usage, e.g., electricity, natural gas or heating oil, into disaggregated end-uses: baseload usage, heating/cooling slope and thermostat setting. PRISM has been used specific applications such as tracking retrofit savings (Mills et al., 1987) or scorekeeping for electricity conversation programs (Dutt and Fels, 1989; Gregory, 1987; Hirst, 1986; Rodberg, 1986). Fels and Reynolds (1992) used PRISM for analyzing New York State Energy Research and Development Authority's (NYSERDA) multifamily conservation program. A similar study was executed by Goldman and Ritschard (1986) for San Francisco Housing Authority to assess energy conservation in public housing. Oak Ridge National Laboratory used PRISM to evaluate national weatherization efforts (Brown et al., 1993). PRISM is not only used for residential buildings or large samples – Haberl et al. (1989) used PRISM to conduct a campus-wide energy performance analysis for Princeton University.

Deriving the savings potential from the difference between the estimated energy efficiency parameter and a reference point, in our case the sample's median, further outlines a framework in our study to statistically predict the savings potential using publically available data (e.g., property tax records and voter registration records), that may arise from different interventions.

Our approach to defining savings potential using a reference points should is only one method among multiple methods encountered in the literature. From a building science perspective, the analyst typically uses engineering formulae that account for detailed structural and thermal characteristics of a given house and its individual components, to compute technical savings potential. The reader can review Urbikain and Sala (2009) for their methods to calculate energy savings related to windows in residential buildings. For further reference, the reader can review Zhu et al. (2009) for their engineering method to calculate savings from new wall construction for two houses in Las Vegas, NV.

Our work is different from the existing PRISM and other load-disaggregation cases in that it uses publically available information on structural and demographic characteristics of houses to statistically predict the PRISM energy-efficiency parameters. This approach overcomes the need for 1) highly private utility data and historical weather data; 2) analytical processing thereof. The underlying motivation is to build statistical models that use only publically available information to generate baseline energy-efficiency intervention potential insight by end-use for all houses in a given service territory. This would electric utilities to design and implement their energy-efficiency programs in an informed fashion. Predicting savings potential by intervention type, e.g., behavioral (smart thermostat) or engineering-based, is helpful for utilities to design their rebate programs effectively and construct personalized messaging campaigns for groups of customers that have savings potential for different end-uses.

Our work additionally investigates how to identify houses that may be subject to a rebound effect. As defined in the literature, a direct rebound effect is the expanded or intensified use of energy arising from efficiency gains and the associated perceived lower cost of energy (Greening et al., 2000; Sorrell et al., 2009; International Risk Governance Council, 2013). For example, it is not uncommon for a homeowner to start setting their thermostat at a higher temperature in the winter, if they have recently undergone an energy efficiency improvement (e.g., attic insulation or replacing the HVAC unit).

Assuming a house would spend realizable savings from energy-efficiency improvements on improvements in their comfort (e.g., adjusting their thermostat setting, until they achieve a level comfort), we attempt to identify the houses that may be subject a rebound effect and quantified the rebound amount.

This exercise is particularly insightful as policy-makers need to measure and verify energy savings, thus ensuring energy efficiency standards and objectives are met. The rebound effect has a significant influence on technically realizable energy efficiency potential and actual realized potential. Extracting the rebound amount as well determining the houses that are most likely to rebound, are valuable when assessing the performance of energy efficiency programs and optimizing their design for prospective efforts.

4.2. Data

PRISM was used to model energy efficiency parameters of single-family houses. Monthly utility usage and historical weather information are primary inputs for PRISM.

PRISM, developed by the Center for Energy and Environmental Studies at Princeton University in 1978, uses daily temperature data from which heating and cooling degree-days are calculated, and monthly utility meter readings for utility consumption as inputs to determine the weather-adjusted index for annual consumption which is called Normalized Annual Consumption (NAC). In essence, NAC is the annual utility consumption for a given year with average weather. In our study we use monthly natural gas usage data as input for the PRISM heating model to compute $NAC_{\text{natural gas}}$, and monthly electricity usage data for the PRISM cooling model to compute $NAC_{\text{electricity}}$

NAC is estimated using a three-parameter model that is a function of thermal integrity of the building (Heating/Cooling Slope – HS/CS), appliance-level baseload consumption (BL), and the interior-temperature setting (Reference Temperature – RT). Baseload natural gas consumption is derived from the heating model and is denoted as $BL_{natural\ gas}$. Similarly, baseload electricity consumption is derived from the cooling model and is denoted as $BL_{electricity}$. Heating slope comes from the heating model as denoted as $HS_{natural\ gas}$ whereas cooling slope comes from the cooling model and is denoted as $CS_{electricity}$. $RT_{natural\ gas}$ and $RT_{electricity}$ come from the heating and cooling models, respectively. Generally, a house's heating/cooling system is operated when the outdoor temperature (T_{out}) goes below/above a certain level (Reference Temperature, RT), and for each incremental degree change in temperature a constant amount of fuel (electricity, fuel oil or natural gas) (the heating/cooling slope (HS/CS) is consumed (Fels, 1986). Hence, the fuel consumed is linearly proportional to ($RT – T_{out}$) and the constant HS/CS represents the house's effective heat-loss (or gain) rate. Further the house may use a constant amount of fuel per day (the base level BL) independent of T_{out}. This is treated as the baseload of the building and is attributed to appliance-level consumption. Thus, PRISM defines the normalized annual consumption for electricity, $NAC_{electricity}$, as:

$$NAC_{electricity} = BL_{electricity} + CS_{electricity} \times CDD(RT_{electricity}) + \varepsilon$$

Where $CDD(RT_{electricity})$ is the number of cooling degree days for a given $RT_{electricity}$ in a given year.

Similarly, normalized annual consumption for natural gas, $NAC_{natural\ gas}$, is defined as:

$$NAC_{natural\ gas} = BL_{natural\ gas} + CS_{natural\ gas} \times HDD(RT_{natural\ gas}) + \varepsilon$$

Where $HDD(RT_{natural\ gas})$ is the number of heating degree days for a given $RT_{natural\ gas}$ in a given year.

The derivation of this equation allows the interpretation of the three parameters: The reference temperature RT, which varies from building to building, is likely to be affected by the indoor temperature T_{in}, which is typically set by a thermostat, and intrinsic gains (e.g., heat produced by appliances and occupants, and the sun). The heat loss (and gain) rate $HS_{natural\ gas}/CS_{electricity}$ is governed by conductive and infiltration heat losses/gains as well as the furnace efficiency (Fels, 1986). The base level consumption BL is determined by the amount of fuel consumed by appliances. ε is the random error term that cannot be explained by the regression equation that is solved by ordinary least-squares linear regression technique. Using an iterative approach based on Newton's method (Goldberg, 1982), PRISM solves for the three parameters, i.e., baseload consumption (BL), heating/cooling slope (HS/CS) and reference temperature (RT), that best explain changes in fuel consumption.

7,091 houses with both electricity and natural gas for 2009-2011 used in the previous chapter were processed by PRISM[12]. Natural gas usage data was used to fit a heating model for each house. For houses with both natural gas and electricity usage, natural gas is primarily used for heating whereas the electricity data was used to fit a cooling model for house since electricity is used for air conditioning. Baseload consumption for natural gas model can be attributed to end-uses like cooking whereas electricity baseload consists mostly of appliance-level consumption, e.g., lighting, refrigerator, TV, etc.

Due to missing or inconsistent utility entries in the original dataset of 7,091 houses PRISM ran successfully only on 5,243 houses. Inconsistent readings can be due to change in occupants or retrofit activities that were not accounted for in the sample..

The percentile distributions of PRISM output parameters, i.e. reference temperature (RT), baseload consumption (BL) and heating/cooling slope (HS/CS) (cooling slope for

[12] For details on the structural and demographic characteristics of houses see Chapter 2.

electricity, heating slope for natural gas usage), normalized annual consumption (NAC), and the associated variances are given in Table 1.

Table 4. 1. Percentiles of PRISM parameters.

	Percentiles				
	5th	**25th**	**50th**	**75th**	**95th**
$R^2_{natural\ gas}$	0.46	0.82	0.88	0.91	0.93
$R^2_{electricity}$	0.23	0.65	0.82	0.90	0.95
$RT_{natural\ gas}$(°C)	11.00	13.00	17.56	19.00	23.63
$RT_{electricity}$(°C)	15.60	19.00	20.92	22.89	25.63
$BL_{natural\ gas}$ (therm/day)	0.12	0.23	0.34	0.47	0.71
$BL_{electricity}$ (kWh/day)	7.55	12.62	17.61	24.70	40.69
$HS_{natural\ gas}$ (therm/degree-day)	0.03	0.21	0.33	0.51	0.93
$CS_{electricity}$ (kWh/degree-day)	1.42	2.50	3.46	4.74	7.80
$NAC_{natural\ gas}$ (therm/year)	115	206	284	379	567
$NAC_{electricity}$ (kWh/year)	4,399	7,380	10,040	13,336	20,176

The median reference temperature is computed as 20.9 degrees for cooling (electricity) and 17.6 degrees for heating (natural gas). The median value for baseload is 17.6kWh/day for electricity and 0.33therms/day for natural gas. Further, median cooling slope (electricity) and heating slope (natural gas) are 3.46kWh/degree-day and 0.33/therms/degree-day, respectively. The median normalized annual consumption as calculated by a combination of reference temperature, baseload and cooling/heating slope, is 284therms/year for natural gas, and 10,040kWh/year for electricity (Table 1). The actual annual usage values for the median house, as given in the Gainesville utility data set, for natural gas and electricity are 284therms/year and 10,068kWh/year. PRISM-predicted annual usage values and actual usage values are comparable which validates the accuracy of PRISM.

4.3. Methods

This study explores the savings potential that may arise from conducting an energy-efficiency intervention among the houses in the sample. The underlying assumption is changing each house's PRISM-computed energy efficiency parameters to the sample's median value for the parameter of interest. Setting the sample median as the reference point is just one approach to examine the energy-efficiency potential of the houses in our sample. Other reference points of interest could be different percentiles (e.g., 5th, 25th, 50th, 75th and 95th) for a given parameter.

,Some houses' energy-efficiency parameters as computed by PRISM, i.e., $RT_{natural\ gas}$, $RT_{electricity}$, $BL_{natural\ gas}$, $BL_{electrcitiy}$, $HS_{natural\ gas}$ and $CS_{electricity}$, are above the median and some are below the median. We define changing reference temperature (RT) values, i.e., adjusting the thermostat setting, as a behavioral intervention and changing baseload consumption (BL) and heating/cooling slope (HS/CS) values as an engineering intervention (e.g., insulation, new refrigerator or new HVAC unit).

A change in consumption would result in savings, i.e., reduced utility usage, if the parameter of interest of for a given house is above the sample's median. Figure 1 illustrates how we calculate change in annual utility consumption (ΔC) for a change in $HS_{natural\ gas}$ for a hypothetical case. The same principle applies for computing ΔC values due to changes in other PRISM output parameters. For a change in reference temperature (RT) the house's comfort level would increase if the $RT_{natural\ gas}$ is increased to the sample's median and $RT_{electricity}$ is reduced to the sample's median. We assumed that an energy-efficiency intervention for BL, HS or CS, is only viable when it reduces the annual utility usage. However, an intervention for RT is considered viable even when the $RT_{natural\ gas}$ is below the sample's median (or when $RT_{electricity}$ is above the sample's median) removing the possibility for savings but creating an opportunity for improved comfort through shifting the RT to the median.

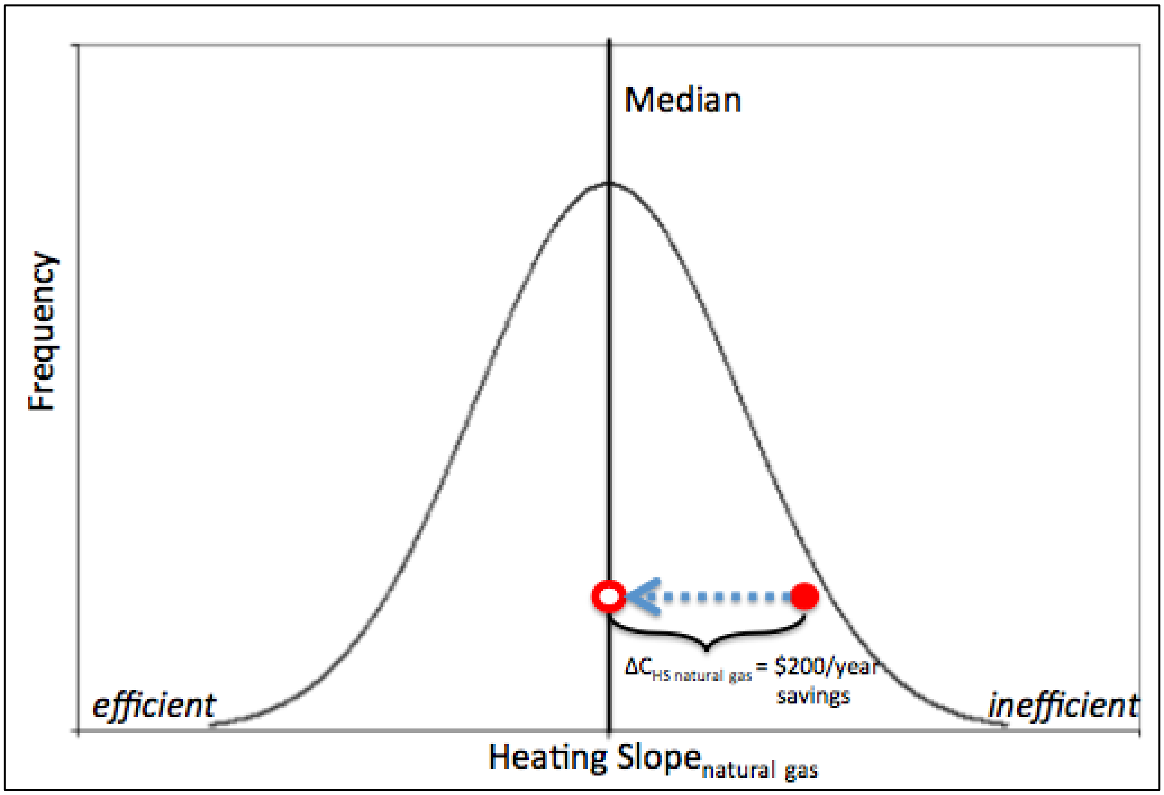
Median
Frequency
$\Delta C_{\text{HS natural gas}}$ = \$200/year savings
efficient
inefficient
Heating Slope$_{\text{natural gas}}$

Figure 4. 1. Procedure to estimate change in consumption (Δ C) because of a change in HS$_{\text{natural gas}}$. The figure shows a hypothetical scenario of \$200/year savings.

We use the PRISM formula (Goldberg and Fels, 1986) to calculate changes in annual utility consumption (ΔC), which are given in dollars per year.

The hypothetical change in annual utility consumption (ΔC) for a change in RT$_{\text{natural gas}}$ to the median is calculated using:

$$\Delta\, C_{\text{RT natural gas}} = \text{Price}_{\text{natural gas}} \times \text{HS}_{\text{natural gas}} \times [\text{HDD}(\text{RT}_{\text{natural gas}}) - \text{HDD}(\text{RT}_{\text{natural gas median}})]$$

Where HDD(RT$_{\text{natural gas}}$) is the number of heating degree days for a given RT$_{\text{natural gas}}$ in a given year.

The $\Delta C_{RT\ natural\ gas}$ is given in dollars per year terms. \$1.35/therms is used as $Price_{natural\ gas}$[13].

Similarly, the hypothetical change in consumption (ΔC) for a change in $RT_{electricity}$ to the median is calculated using:

$$\Delta\ C_{RT\ electricity} = Price_{electricity} \times CS_{electricity} \times [CDD(RT_{electricity\ median}) - CDD(RT_{natural\ gas})]$$

Where $CDD(RT_{electricity})$ is the number of cooling degree days for a given $RT_{electricity}$ in a given year.

\$0.115/kWh was used as $Price_{electricity}$[13].

$HS_{natural\ gas}$ and $CS_{electricity}$ are a function of the building shell's lossiness and the thermal efficiency of the heating or cooling system, respectively. Given that larger buildings with larger shells inherently tend to have larger $HS_{natural\ gas}$ and $CS_{electricity}$ we normalized $HS_{natural\ gas}$ and $CS_{electricity}$ with respect to the building's square footage (SQFT) by dividing each house's $HS_{natural\ gas}$ and $CS_{electricity}$ by its SQFT to compute HS_{SQFT} and CS_{SQFT}.

Thus, the hypothetical change in consumption (ΔC) from a change $HS_{natural\ gas}$ to the median is computed using:

$$\Delta\ C_{HS\ natural\ gas} = Price_{natural\ gas} \times HDD(RT_{natural\ gas}) \times SQFT \times (HS_{SQFT} - HS_{SQFT\ median})$$

Similarly, the hypothetical change in consumption (ΔC) from a change $CS_{natural\ gas}$ to the median is computed using:

$$\Delta\ C_{CS\ electricity} = Price_{electricity} \times CDD(RT_{electricity}) \times SQFT \times (CS_{SQFT} - CS_{SQFT\ median})$$

The hypothetical change in consumption (ΔC) that can realize from changing the $BL_{natural\ gas}$ to the median is determined using:

[13] Average natural gas and electricity retail prices for residential users in 2010 were used as retrieved from the U.S. Department of Energy's Energy Information Administration.

$$\Delta\ C_{\text{BL natural gas}} = 365 \times \text{Price}_{\text{natural gas}} \times (\text{BL}_{\text{natural gas}} - \text{BL}_{\text{natural gas median}})$$

Similary, the hypothetical change in consumption (ΔC) that can realize from changing the $\text{BL}_{\text{electricity}}$ to the median is determined using:

$$\Delta\ C_{\text{BL electricity}} = 365 \times \text{Price}_{\text{electricity}} \times (\text{BL}_{\text{electricity}} - \text{BL}_{\text{electricity median}})$$

Tables 2 and 3 show the percentiles of ΔC potential in dollars per year arising from changing individual houses' PRISM parameters to the sample's median. Ostensibly, some houses' PRISM values are more "efficient" than the sample's median, which results in negative values. Positive ΔC values imply that the respective energy-efficiency parameter is above the sample's median whereas a negative ΔC value implies that the respective energy-efficiency parameter is below the sample's median.

Table 4. 2. Percentiles of ΔC potential in dollars per year arising from changing individual houses' PRISM parameters to the sample's median.

			Percentiles		
ΔC($/year)	**5th**	**25th**	**50th**	**75th**	**95th**
BL-NG	6	29	65	115	233
BL-E	22	136	298	587	1,251
HS	6	28	61	114	247
CS	7	38	82	161	329

Table 4. 3. Percentiles of ΔC potential in dollars per year arising from changing individual houses' reference temperature parameters to the sample's median. Positive values indicate a savings potential whereas negative values indicate a comfort improvement potential.

			Percentiles		
ΔC($/year)	**5th**	**25th**	**50th**	**75th**	**95th**
RT-NG	-535	-210	0	41	160
RT-E	-402	-144	0	136	489

It was assumed a positive ΔC from an intervention to improve $HS_{natural\ gas}$ ($CS_{electricity}$), (e.g., insulation or replacing the HVAC unit), which implies that $HS_{natural\ gas}$ ($CS_{electricity}$) is above the sample's median, would lead the occupants, whose $RT_{natural\ gas}$ ($RT_{electricity}$) is lower (higher) than the sample's median, to seek improved comfort via increasing (decreasing) their thermostat setting during heating (cooling) season. In other words, technical efficiency potential that may arise from an engineering intervention, may not be fully realized as some occupants may choose to spend the potential savings on improving their comfort via setting their thermostat higher for winter and lower for summer. This phenomenon, also known as the rebound effect, hinders complete savings realization. For this, we explored the savings distribution from improvements in heating and cooling slope versus reference temperatures. Figures 2 and 3 show the ΔC distributions arising from changing the reference temperature versus cooling slope (or heating slope). This gives a context on how changing the heating or cooling slope compare to changes in thermostat setting in terms of potential change in annual utility usage.

The first quadrant represents the houses where the thermal integrity is higher, i.e., less efficient, than the sample's median, resulting in a positive ΔC potential. Further, the $RT_{natural\ gas}$ ($RT_{electricity}$) of these houses are above (below) the sample's median, creating additional savings potential with a change in thermostat setting, i.e., decreasing $RT_{natural\ gas}$ and increasing $RT_{electricity}$.

The second quadrant in both figures shows the houses where the occupants live in relatively "comfortable" thermostat setting, defined as being below the sample's median for cooling thermostat setting ($RT_{electricity}$) and above the sample's median for heating thermostat setting ($RT_{natural\ gas}$), generating savings potential for a change in thermostat setting (i.e., increasing $RT_{electricity}$ and decreasing $RT_{natural\ gas}$). However, these houses' thermal integrity ($CS_{electricity}$ and $HS_{natural\ gas}$) is already lower, i.e., more efficient, than the sample's median making them unlikely candidates for an intervention, e.g., insulation.

The third quadrant in both figures, denote the houses where the thermal integrity, i.e. heating or cooling slope, is lower, i.e., more efficient, than the median. The same houses live in an already below median thermostat setting for heating and above median for cooling, thus unlikely to generate any potential for savings from changing the thermostat setting. Similar to the fourth quadrant, the houses herein have comfort improvement potential where the dollar amount denotes how much additional money the house would have to spend on adjusting the thermostat setting to move to the median.The fourth quadrant contains the houses with a potential rebound effect. In other words, the savings potential from improving the thermal integrity is positive whereas the thermostat is set at a relatively uncomfortable level, i.e., $RT_{\text{natural gas}}$ is below the sample's median and $RT_{\text{electricity}}$ is above the sample's median. This deviation from the median reference temperature implies a potential improvement in social welfare as the ambient temperature in a house affects the comfort level, air quality and occupants' health. It is assumed that the occupants of a house will spend the savings gained from reduced utility bills because of thermal integrity improvements, to change its thermostat setting until it reaches the median thermostat setting. We assume that people who live in houses that have thermostat setting that are below (above) the sample's median in the winter (summer) would prefer to set their thermostats at a warmer (cooler) temperature if it was affordable. Some houses could realize enough savings from thermal integrity improvement to change their thermostat towards the median whereas others could see a partial comfort improvement because of limited savings from thermal integrity interventions.

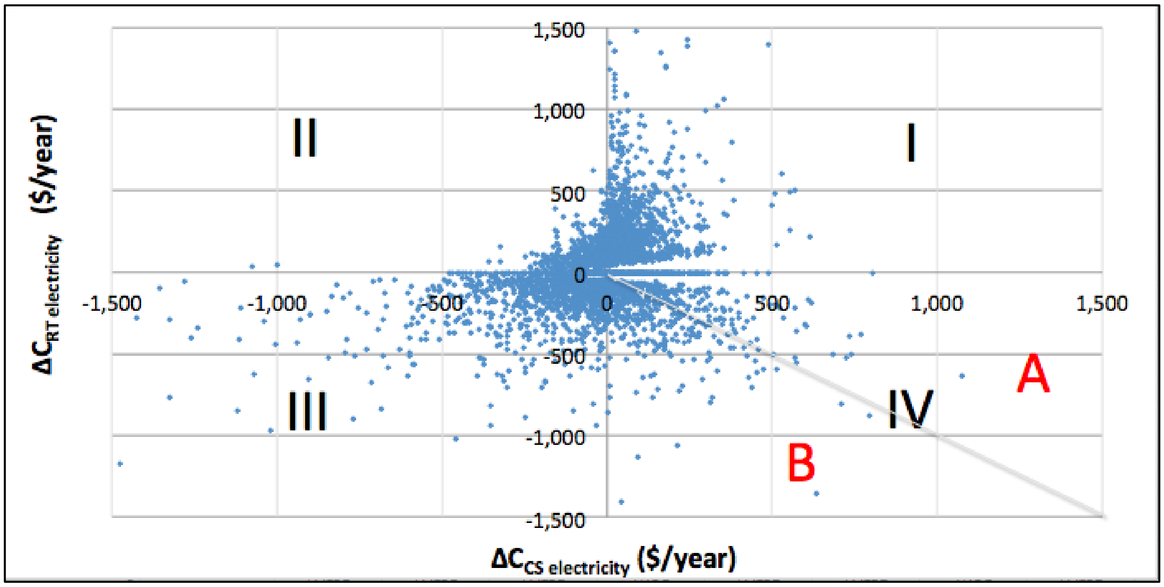

Figure 4. 2. ΔC from change in reference temperature for versus ΔC from change in cooling slope (electricity).

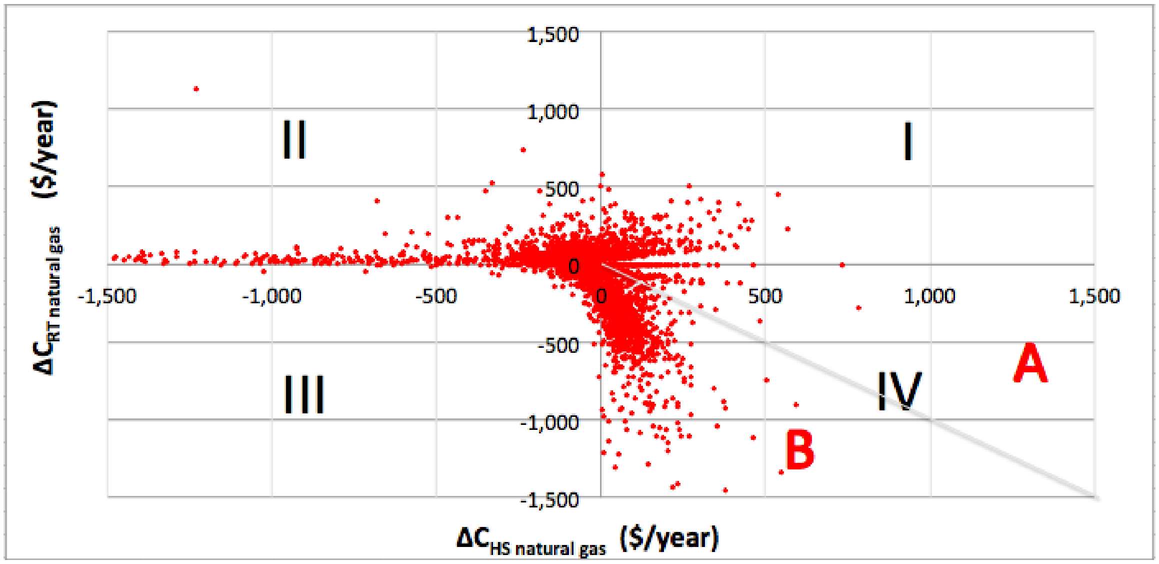

Figure 4. 3. ΔC from change in reference temperature for versus ΔC from change in heating slope (natural gas).

Table 4 shows the breakdown of the four quadrants for electricity consumption models. Most houses fall in quadrants 1 (1,428 houses) and 3 (1,518 houses). Quadrant 2 with efficient cooling slopes and lower higher than median $RT_{electricity}$, has the lowest number of houses. The rightmost column in Table 4 shows the aggregate potential across the four quadrants. Aggregate savings from increasing the thermostat setting ($RT_{electricity}$)

pose approximately a $1.2M/year potential whereas to bring houses below the median thermostat setting to the median, an investment of 0.5M/year would be necessary to spend on this comfort improvement. Further the aggregate savings potential on the thermal integrity improvement front amounts to approximately $260K/year for the sample. Interestingly, if energy efficiency interventions were conducted across the sample this would result in potential savings of approximately $1.5M/year which is larger than the amount it would take to move the houses above median reference temperature to the median ($0.5M/year) (Table 4).

Table 4. 4. Breakdown of four quadrants in Figure 1 for electricity consumption. The first three rows show the sum of savings or increased comfort potential for each quadrant.

	Change in Consumption ($/year)				
# of houses	1,428	761	1,518	855	4,562
($/year)	**Quadrant 1**	**Quadrant 2**	**Quadrant 3**	**Quadrant 4**	**Sum**
$RT_{savings}$	1,095,199	96,912	-	-	1,192,111
$RT_{comfort}$	-	-	-298,979	-199,655	-498,634
$CS_{savings}$	131,152	-	-	129,241	260,393

Similarly, Table 5 shows the breakdown of four quadrants for natural gas consumption models. The highest number of houses fall into quadrants 2 (1,377 houses) and 4 (1,817 houses). Quadrant 3 has the lowest number of houses. The rightmost column in Table 4 shows the aggregate potential across the four quadrants. The aggregate savings potential from decreasing the thermostat setting ($RT_{natural\ gas}$) totals to $171K/year for our sample. Further, an investment of $730K/year would be necessary to provide a median level comfort to the houses, which are below this level. The collective savings potential, which may arise from improving the houses to the median level thermal integrity, amounts to $197K/year. Unlike energy efficiency interventions for electricity, the aggregate savings potential including thermostat adjustment and heating slope improvements amounts to approximately $367K which is smaller than what is needed to move the houses to the median reference temperature ($730K) (Table 5).

Table 4. 5. Breakdown of four quadrants in Figure 2 for natural gas consumption. The first three rows show the sum of savings or increased comfort potential for each quadrant.

	Change in Consumption ($/year)				
# of houses	514	1,377	757	1,817	4,465
($/year)	**Quadrant 1**	**Quadrant 2**	**Quadrant 3**	**Quadrant 4**	**Sum**
$RT_{savings}$	70,531	99,935	-	-	170,466
$RT_{comfort}$	-	-	-77,158	-652,713	-729,871
$HS_{savings}$	58,110	-	-	138,808	196,918

The assumption was made that whenever a house improves its thermal integrity through an intervention (i.e. insulation, HVAC unit replacement), to the sample's median value, the potential savings will be spent towards adjusting the thermostat setting till the comfort level reaches the sample's median reference temperature. In other words, technical energy efficiency potential will not materialize until a given house is median-level comfortable in terms of thermostat setting. Ostensibly, some houses have large enough thermal integrity improvement potential that some saving potential is left for comfort improvement ($\Delta C_{HS} > |\Delta C_{RT}|$) (region A above the diagonal in Figures 2 and 3) and some do not ($\Delta C_{HS} < |\Delta C_{RT}|$) (region B below the diagonal in Figures 2 and 3). The breakdown of these rebound houses, which are also shown in the fourth quadrant in Figures 2 and 3, is given in Table 6.

Table 4. 6. Breakdown of houses with potential rebound effect. "A" denotes the region in quadrant 3 above the diagonal ($\Delta C_{HS\ or\ CS} > |\Delta C_{RT}|$) and "B" denotes the region in quadrant 4 below the diagonal ($\Delta C_{HS\ or\ CS} < |\Delta C_{RT}|$) in Figures 1 (cooling) and 2 (heating).

	Change in Consumption ($/year)				
	Heating		**Cooling**		**Sum**
	A	**B**	**A**	**B**	
$RT_{comfort}$	-11,965	118,390	-37,810	69,566	138,181
$HS/CS_{savings}$	20,418	-	59,676	-	80,094
# of houses	130	1,687	222	633	2,672

In order to put the savings potential from a thermostat adjustment or a thermal integrity improvement in perspective, the net present value (NPV) of annual savings of $100, $200, $300, $400, $500, and $600 over 10 years at 0%, 5%, 10% and 20% discount rates is calculated. This range of hypothetical annual savings can be extended. However, we used $0-$600 since that covers 99%+ of our sample's savings potential at an individual house level (Table 7). This benefit-cost context illustrates that what kind of annual savings at a given discount rate could allow a certain intervention. For example, a hypothetical $300/year savings at 5% generates an NPV of $2,317 and could accommodate a furnace replacement which can cost $1,450 (Table 8). This NPV-based approach can be particularly insightful for utility energy efficiency program managers when allocating rebate amounts to specific energy efficiency measures as well as when designing subsidized loan programs. Table 9 shows the costs of different energy efficiency measures that can be implemented.

Table 4. *7*. Savings potential and number of houses by individual energy efficiency parameter changes.

Savings potential	**# of houses**					
($/year)	**RT-NG**	**RT-E**	**BL-NG**	**BL-E**	**HS**	**CS**
0	3,351	3,053	2,555	2,647	2,622	2,622
1-100	1,301	611	1,842	485	1,836	1,512
101-200	423	648	633	449	565	660
201-300	110	380	151	370	150	273
301-400	36	185	37	278	36	104
401-500	12	115	18	232	19	36
501-600	7	64	3	153	10	18

Table 4. 8. Net present values (NPV) of hypothetical annual savings amounts ($100-$600) across different discount rates (0%-20%).

	Annual Savings ($)					
Discount rate (%)	**100**	**200**	**300**	**400**	**500**	**600**
0	$1,000	$2,000	$3,000	$4,000	$5,000	$6,000
5	$772	$1,544	$2,317	$3,089	$3,861	$4,633
15	$502	$1,004	$1,506	$2,008	$2,509	$3,011
20	$419	$838	$1,258	$1,677	$2,096	$2,515

Table 4. 9. Costs of potential energy-efficiency measures[14]

High efficiency central AC	$1,450
High efficiency furnace	$1,100
Attic insulation	$1-2/sqft
Wall Insulation	$0.75/sqft
High efficiency refrigerator	$2,000

As described previously heating or cooling slope of a house is a function of the lossiness of the building shell as well as the efficiency of the heating or cooling unit. Looking at heating or cooling slope alone does not allow for determining whether the house has an insulation problem or needs a heating or cooling unit replacement or all three. However, plotting normalized values of heating and cooling slopes (Figure 4) generates some insight as to which houses have a heating or a cooling unit problem. In other words, if the normalized value of heating slope is large (i.e., inefficient) when the normalized value of cooling slope is small (i.e., efficient), it is likely that the heating unit should be replaced (shown in the second quadrant in Figure 4). The same principle applies for a cooling unit replacement as illustrated in quadrant four in Figure 4. Quadrant one in Figure shows the houses that have high normalized values for both heating and cooling slopes; they have poor insulation and inefficient heating and cooling units. Moreover, quadrant three depicts the houses with low normalized values for heating and cooling slopes, which imply good insulation and good efficiency of heating and cooling units. Processing natural gas and electricity usage as described in our framework facilitates a comparison between heating and cooling profiles and provide insight on identifying which houses fall into which quadrants. Unfortunately, additional data input is necessary to segregate between insulation or heating/cooling unit efficiency issues for quadrants one and three (Figure 4). Table 10 shows the numbers of houses in our sample that fall into different quadrants in Figure 4. It is implied that 447 (second quadrant in Figure 4 and Table 10) houses could benefit from replacing their heating unit whereas 1,323 houses could benefit from replacing their cooling unit (fourth quadrant in Figure 4 and Table 10).

[14] The prices were gathered from Gainesville-area vendors and contractors via phone.

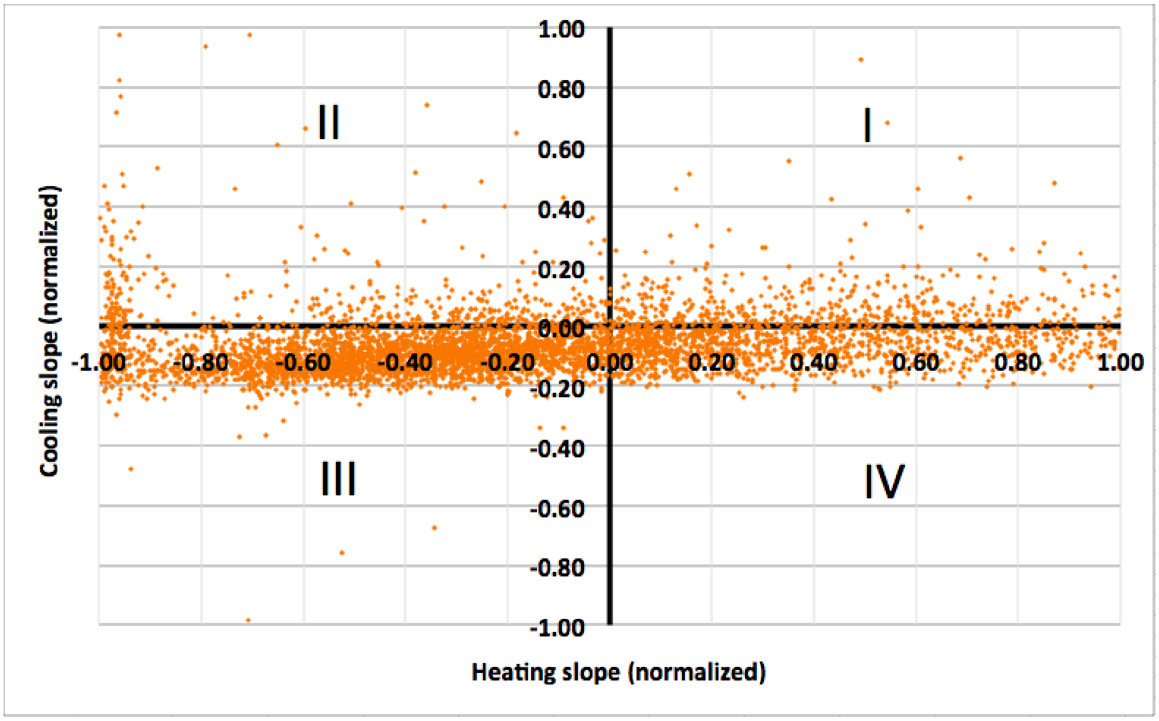

Figure 4. 4. Heating slope normalized versus Cooling slope normalized.

Table 4. 10. Number of houses in four quadrants in Figure 3.

Quadrant	# of houses
I	680
II	447
III	2,793
IV	1,323

4.4. Models

Weighted least squares regression was used to predict the change in utility usage potential (ΔC in $/year) that can arise from changing the PRISM efficiency parameters, i.e., reference temperature, heating/cooling slope, baseload consumption, to the sample's median values using structural and demographic house characteristics as explanatory variables. The previous chapter discusses the advantages of weighted least squares regression method for when each observation's variability can be measured. Each data point represents a PRISM simulation and how it relates to the sample's median.

Table 11 provides details for the ΔC regression models for reference temperature (RT), baseload (BL), heating/cooling slope (HS/CS) and rebound amount both for electricity and natural gas, and the resulting statistically significant independent variables ($p=0.05$).

The resulting R^2's vary between 6% and 95%, Reference Temperature (RT) for the heating/natural gas model having the lowest (5%) and cooling slope electricity savings potential having the highest (95%). The analyst can further used predicted ΔC values to compute rebound potential on natural gas and electricity usage for each house.

Table 4. 11. Regression coefficients for ΔC models in dollars per year. The coefficients are statistically significant at p=0.05.

Independent Variables				Models					
	Minimum	Median	Maximum	$RT_{natural\ gas}$	$RT_{electricity}$	$HS_{natural\ gas}$	$CS_{electricity}$	$BL_{natural\ gas}$	$BL_{electricity}$
Land value ($1000s)	0	30	479.7	0.21	-108	0.23	-4.83	-0.18	301
Building value ($1000s)	17.4	106.1	664.9	0.12			-3.72	0.56	-985
Misc. value ($1000s)	0	1.8	28.4	0.43	-107	1.06	12.33	0.68	-2,578
Tax amount ($1000s)	0	1.8	17.1				281.75	18.50	-3,629
Actual area (1000sqft)	0.6	2.2	10.3	-8.97	5,541		54.26	-45.73	76,537
Heated area (1000sqft)	0.5	1.6	7.5	-18.49	-4,139	-16.11	-654.60	-25.07	-25,838
Age of the building	4	33	70	-0.40	131		2.01	-1.40	582
# of bedrooms	1	3	5	4.67	-2,212	6.37	29.88	12.70	48,369
# of bathrooms	1	2	7	10.70	4,180	6.56	240.98	-37.88	25,285
# of stories	0	1	3	11.91	788	-8.49	237.12	4.74	-20,474
Avg. age of occupants	18.2	51.8	90			-0.16		0.75	-1,632
Avg. years of occupancy	0	15	87	0.57			-4.67	0.59	1,852
Total # of occupants	1	2	7			5.52	-83.49	90.98	-64,101
# of teenagers	0	0	2			9.06	188.03	99.62	-39,683
# of seniors	0	1	4	5.01		-2.32	97.35	-5.42	5,647
# of republicans	0	0	5	-7.01	-1,535	3.17	121.27	2.73	-13,796
# of democrats	0	1	7	-5.74	-1,731	3.22	82.95	19.07	
# of males	0	1	5					-73.55	27,790
# of females	0	1	7			-6.38		-96.40	73,074
Intercept					-5,113			-79.34	-18,267
R^2				0.06	0.09	0.47	0.95	0.54	0.89
n				5,117	5,110	5,118	5,113	5,243	5,243

4.5. Discussion

This study illustrates that processing monthly utility billing data and historical weather data by a load disaggregation software like PRISM can produce insightful results in creating the efficiency profile of houses in a given service territory. Understanding the breakdown of utility consumption, e.g. baseload consumption, heating/cooling-related usage and reference temperature, can be helpful for comparing individual houses reference points of interest, e.g., a sample's median. We chose the median value as an improvement threshold for a demonstration but other reference points can be used to explore viable energy-efficiency candidates. This approach also allows categorizing houses into subgroups for a targeted energy-efficiency implementation strategy, i.e., houses with thermal-integrity improvement potential; with comfort improvement potential; with thermostat setting adjustment potential; as well as rebound potential.

Assessing a large number of houses on the different aspect of their efficiency profile can be powerful when implementing utility energy-efficiency programs. Computing Net Present Value (NPV) of annual savings potential across the different energy efficiency parameters can further shed insight on what kind of interventions may be deemed economically attractive from a residential user's perspective as well as what kind of additional financial incentives, e.g. rebates and low-interest loan programs, would be necessary to shift customers' energy efficiency appetite from one bracket to another from an economic standpoint.

The overarching objective of our work was to categorize houses based on energy-efficiency characteristics and demonstrate that using a simple data-driven approach may generate value in execution of analytically targeted energy efficiency program outreach. The energy-efficiency decision and value chains consist of multiple parts, from a home audit to the actual deployment, where multiple stakeholders partake. Before a costly home audit, the stakeholders (e.g., utility, product and services providers) ought to diagnose what energy-efficiency issues a house may have in a non-invasive and data-driven fashion. This in turn will enable stakeholders to allocate the right marketing resources and use appropriate customer messaging tools to increase customer conversion and thus the return on their marketing dollars.

Given the relatively significant R^2s of our weighted least squares regression models, publically available information can be helpful in assessing residential building stocks' savings potential on different intervention types further improving targeted customer outreach.

The reader should note that PRISM is designed to compute normalized annual consumptions (NAC) and the individual parameters that make are more uncertain (Fels, 1986). Although PRISM calculates the standard error associated with each output parameter, it is hard to interpret the individual PRISM parameters for a given house because of the uncertainty. This is reflected on our regression models whose beta coefficients are derived in part from the uncertain PRISM parameters.

4.6. Conclusion

The ecosystem is subject to many market failures and barriers and the absence of access to highly private utility usage data further exacerbates the situation. Our framework builds predictive models based on publically available data to overcome this problem. The generated models can be used by EE program implementers who may not have access to digitally available utility usage information, to design, implement and evaluate their programs. Assessing the savings potential by end-use as proposed by our work can facilitate effective EE program outreach and messaging so that the EE program implementers can allocate the right resources to the right consumers for a given EE "reservoir".

Our study proposed an easy-to-execute framework on disaggregating utility usage information into actionable end-use insight. We believe that with the advent of smart grid, big data, and improved cloud computing, electric utilities will become savvier and data-driven in their energy-efficiency program implementation. Our framework illustrates that a monthly-bill-based approach can be used in load disaggregation, which can be extended into exploring the distributions in individual energy efficiency parameters. Once the regression models are built, energy-efficiency intervention estimates can be determined with only publicly available information. Customer-specific outreach and messaging is necessary in compelling consumers to become more energy efficient and can be facilitated using our framework based on publically available data allowing multiple stakeholders to influence the market.

Innovation Electricity Efficiency (IEE), an institute of the Edison Foundation, forecasts that utility energy-efficiency program expenditures can increase from its current national aggregate level of $6.9B to $14.3B by 2025 (2013). While the energy-efficiency programs provide financial incentives and are critical, other stakeholders (e.g., home improvement product and service providers, contractors and lenders) play a pivotal role as well. In the absence of highly private utility usage data non-utility stakeholders experience significant limitations to making strategic decisions in a data-driven and analytical fashion. This hinders energy-efficiency marketing outreach and deployment. Our work based on regression models using publically available data reduces this barrier allowing energy-efficiency stakeholders informed decisions.

Utilities ought to understand their customer-base better at individual house and potential intervention prospect level. Future research should examine the validity of our estimates and conduct this exercise using higher frequency utility data, e.g., hourly usage, which may bear additional value. Moreover, the validity of our assumptions and results regarding the regression analyses should be tested by extending this work to other geographic regions since the statistical predictions accuracy may be undermined by where the publicly available data come from.

There are significant limitations to our work. Voter registration records have information only on the registered citizens who are eligible to vote, i.e., no information on non-voters and children. Further, the models underlie PRISM-processed data whose results are uncertain. PRISM does successfully process house data with missing or inconsistent utility readings – we had to remove these houses from our end sample (n=5,243). This can result in substantial bias in our models.

4.7. References

1. American Council for an Energy-Efficient Economy, "State Energy Efficiency Resource Standard (EERS) Activity". Washington, D.C. 2011
2. M. Brown, L. Berry and L. Kinney, "Weatherization Works: An Interim Report of the National Weatherization Evaluation". Report No. ORNL/CON-373, Oak Ridge National Laboratory, Oak Ridge, TN. 1993
3. W. Chung, "Review of Building Energy-use Performance Benchmarking Methodologies". *Applied Energy*. 2011, Vol. 88, pp. 1470-1479
4. G. Dutt and M. Fels, "Keeping Score in Electricity Conservation Programs. Electricity: Efficient End-Use and New Generation Technologies, and Their Planning Implications". Vattenfall Electricity Congress. Lund University Press, Lund, Sweden. 1989, pp. 353- 388
5. K. Ehrhardt-Martinez, A. K. Donnelly and J. A. Laitner, "Advanced Metering Initiatives and Residential Feedback Programs: A Meta-Review for Household Electricity-Saving Opportunities". Report Number E105. American Council for an Energy-Efficient Economy, Washington, D.C. 2010
6. L. A. Greening, D. L. Greene and C. Difiglio, "Energy Efficiency and Consumption – the Rebound Effect – A Survey". *Energy Policy*. 2000, Vol. 28, pp. 389-401
7. M. F. Fels, "PRISM: An Introduction". Center for Energy and Environmental Studies, Princeton University, Princeton, NJ 08544. 1986
8. M. Fels and C. Reynolds, "Energy Analysis in New York City Multifamily Building: Making Good Use of Available Data". Report NYSERDA 93-3, New York State Energy Research and Development Authority, Albany, NY. 1992
9. C. Goldman and R. Ritschard, "Energy Conservation in Public Housing: a Case Study of the San Francisco Housing Authority". *Energy and Buildings*. 1986, Vol. 9, pp. 89-98.
10. J. Gregory, "Ohio Home Weatherization Assistance Program Final Report". Office of Weatherization, Ohio Dept. of Development, Columbus, OH. 1987

11. J. S. Haberl, S. Englander, C. Reynolds, M. McKay and T. Nyquist, “Whole-Campus Performance Analysis Methods: Early Results from Studies at the Princeton Campus.” Proceedings of 6th Annual Symposium on Improving Energy Efficiency in Hot and Humid Climates. Texas A & M University, Dallas, TX. 1989
12. E. Hirst, “Electricity Savings One, Two and Three Years After Participation in the BPA Residential Weatherization Pilot Program”. *Energy and Buildings*. 1986, Vol. 9, pp. 45-53.
13. Institute for Electric Efficiency, The Edison Foundation, “Summary of Ratepayer-Funded Electric Efficiency Impacts, Budgets, and Expenditures”. 2012
14. Institute Efficiency Innovation. “Summary of Customer-Funded Electric Efficiency Savings, Budgets, and Expenditures (2011-2012)”. 2013
15. International Risk Governance Council, “ The Rebound Effect: Implications of Consumer Behavior for Robust Energy Policies”. ISBN 978-2-9700772-4-4. Lausanne, Switzerland. 2013
16. A. B. Jaffe and R. N. Stavins, “Energy-Efficiency Investments and Public Policy”. *The Energy Journal*. 1994, Vol. 15, No.2.
17. A. Kavousian and R. Rajagopal, “Data-Driven Benchmarking of Building Energy Efficiency Utilizing Statistical Frontier Models”. *Journal of Computing in Civil Engineering*. 2013. 10.1061/(ASCE)CP.1943-5487.0000327
18. E. Mills, M. Fels and C. Reynolds, “PRISM: A Tool for Tracking Retrofit Savings. *Energy Auditor and Retrofitter*”. 1986, Nov/Dec., pp. 27-34.
19. National Academy of Sciences, “Real Prospects for Energy Efficiency in the United States”. The National Academies Press, Washington, D.C. 2010
20. L. Rodberg, “Energy Conservation in Low-Income Homes in New York City: The Effectiveness of House Doctoring”. *Energy and Buildings*. 1986, Vol. 9, pp. 55-64.
21. S. Sorrell, J. Dimitropoulos and M. Sommerville, M. “Empirical Estimates of the Direct Rebound Effect: A Review”. *Energy Policy*. 2009, Vol. 37, pp. 1356-1371.
22. M. K. Urbikain and J. M. Sala, “Analysis of Different Models to Estimate Energy Savings Related to Windows in Residential Buildings”. *Energy and Buildings*. 2009, Vol. 41, pp. 687-695.

23. L. Zhu, R. Hurt, D. Correia and R. Boehm, “Detailed Energy Saving Performance Analyses on Thermal Mass Walls Demonstrated in a Zero Energy House”. *Energy and Buildings*. 2009, Vol. 41, pp. 303-310.

5. Conclusions and Policy Implications

In our work we have shown that publically available data can be used to model residential utility usage in the absence of highly private utility data. We have demonstrated that accounting for billing lag when attributing usage to individual months is of critical importance, particularly for statistical modeling of monthly usage. The accuracy of such models can be diminished if billing lag is not addressed as it can cause significant deviations.

We acknowledge that our approach has limitations whose extent remains to be explored. In other words, it is unknown how accurately our models based on the Gainesville sample can predict utility usage for houses outside that are outside our sample, e.g., other houses in Gainesville, Florida or other regions. Collection of further data in other locations can shed light on the extent of our models' accuracy. Comparing these results to Department of Energy's Residential Energy Consumption Survey (RECS), which harbors aggregate residential energy data for different regions in the US., can generate further insight on our accuracy as well as provide suggestions on how better RECS can be designed and collected. This prospective sample/out-of-sample analysis can prove useful in understanding specific limitations of our models and the independent variables of interest.

Additionally, we did not incorporate any demographic information on occupants who are not eligible and registered to vote. Our sample did not account for non-voters and children which can cause significant bias. Also, PRISM does not process house data with missing or inconsistent utility readings. We had to remove these houses from our sample and this can cause additional bias in our final models.

The regression models built have significant explanatory power in illustrating the utility usage that can be used by policy makers and third-party developers and operators engaged in energy efficiency and real estate businesses for strategic planning. Aggregate data like RECS does not allow for profiling individual houses and using statistical models based on publically available data can establish the first step to examine the geographic variations in utility usage for large regions. This will pave a path to study what physical and demographic factors drive usage and how they should be treated to promote energy efficiency deployment. This can help utilities plan for future energy demand and power generation capacity.

Processing utility usage information using software like PRISM can help building baselines and define efficiency profiles for individual houses in utility service territories. However, utility usage information is highly private and typically not shared with 3rd parties. This work shows that such publically available information can help predict efficiency parameters, as computed by PRISM, i.e., thermostat setting, thermal integrity, baseload consumption, within varying degrees of explanatory power. Our findings further underscore that availability of data and analytical use thereof are critical for understanding the US building stock for energy efficiency targeting and accelerate energy efficiency deployment. Expanding available datasets to different frontiers such as smart meters, smart thermostats or house audits can strengthen the potential analytical insight we can derive.

Being able to use publically available data and accurately predict structure thermal integrity, baseload consumption and thermostat setting are valuable because every structure has different physical and demographic characteristics and only knowing what constitutes an energy-efficiency problem can lead to formulation of an intelligent intervention. Only disaggregating utility usage into different end-uses can facilitate diagnosing residential energy issues. Using statistical models based on publically available data on individual houses allows this diagnostic exercise to scale up for large regions which provides critical input for utilities and policy-makers to develop analytically-driven energy-efficiency targeting strategies. This in turn will enable decision-makers to assess residential building stocks from an economic as well as a social welfare perspective, and design, implement and evaluate their energy-efficiency programs systematically. Utilization of publically available data and statistical models, as proposed by our work, can overcome the barrier to access to highly private utility usage data and can help EE program implementers allocate the right resources to the right consumers and EE measures to meet their state-mandated demand-reduction targets in an analytical and informed fashion.

Our study proposed an easy-to-execute framework on disaggregating utility usage information into actionable end-use insight. We believe that with the advent of smart grid, big data, and improved cloud computing, electric utilities will become savvier and data-driven in their energy-efficiency program implementation. Our framework illustrates that a monthly-bill-based approach can be used in load disaggregation, which can be extended into exploring the distributions in individual energy efficiency parameters. Once the regression models are built, energy-efficiency intervention estimates can be determined with only publicly available information. Customer-specific outreach and messaging is necessary in compelling consumers to become more energy efficient and can be facilitated using our framework based on publically available data allowing multiple stakeholders to influence the market.

Innovation Electricity Efficiency (IEE), an institute of the Edison Foundation, forecasts that utility energy-efficiency program expenditures can increase from its current national aggregate level of $6.9B to $14.3B by 2025 (2013). While the energy-efficiency programs provide financial incentives and are critical, other stakeholders (e.g., home improvement product and service providers, contractors and lenders) play a pivotal role as well. In the absence of highly private utility usage data non-utility stakeholders experience significant limitations to making strategic decisions in a data-driven and analytical fashion. This hinders energy-efficiency marketing outreach and deployment. Our work based on regression models using publically available data reduces this barrier allowing energy-efficiency stakeholders informed decisions.

Assessing a large number of houses on the different aspect of their efficiency profile can be powerful when implementing utility energy-efficiency programs. Computing Net

Present Value (NPV) of annual savings potential across the different energy efficiency parameters can further shed insight on what kind of interventions may be deemed economically attractive from a residential user's perspective as well as what kind of additional financial incentives, e.g., rebates and low-interest loan programs, would be necessary to shift customers' energy efficiency appetite from one bracket to another from an economic standpoint.

The energy-efficiency decision and value chains consist of multiple parts, from a home audit to the actual deployment, where multiple stakeholders partake. Before a costly home audit, the stakeholders (e.g., utility, product and services providers) ought to diagnose what energy-efficiency issues a house may have in a non-invasive and data-driven fashion. This in turn will enable stakeholders to allocate the right marketing resources and use appropriate customer messaging tools to increase customer conversion and thus the return on their marketing dollars.

Utilities ought to understand their customer-base better at individual house and potential intervention prospect level. Future research should examine the validity of our estimates and conduct this exercise using higher frequency utility data, e.g., hourly usage, which may bear additional value. Moreover, the validity of our assumptions and results regarding the regression analyses should be tested by extending this work to other geographic regions since the statistical predictions accuracy may be undermined by where the publicly available data come from.