Title
Targeted Efficiency: Using Customer Meter Data to Improve Efficiency Program Outcomes

Author
Borgeson, Samuel Dalton

Publication Date
2013

Peer reviewed|Thesis/dissertation

Targeted Efficiency: Using Customer Meter Data to Improve Efficiency Program Outcomes

by

Samuel Dalton Borgeson

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Energy and Resources

in the

Graduate Division
of the
University of California, Berkeley

Committee in charge:

Professor Duncan Callaway, Chair
Professor Daniel Kammen
Professor Ram Rajagopol
Professor Catherine Wolfram

Fall 2013

Targeted Efficiency: Using Customer Meter Data to Improve Efficiency Program Outcomes

Abstract

Targeted Efficiency: Using Customer Meter Data to Improve Efficiency Program Outcomes

by

Samuel Dalton Borgeson
Doctor of Philosophy in Energy and Resources

University of California, Berkeley

Professor Duncan Callaway, Chair

Energy efficiency (EE) and demand response (DR) programs are designed to reduce energy consumption, mitigate grid capacity constraints, support intermittent renewable energy integration, and reduce pollution for less than the cost of additional generation. However, the savings and flexibility achieved by EE and DR programs are contingent on programs finding and enrolling customers well matched to program objectives. Many EE programs currently rely on broad but shallow savings from prescriptive measures, but growing interest in deeper and more reliable savings is leading to increased attention for program planning and targeting using empirical criteria. Through data gathered by smart meters and software designed to manage and analyze large data sets, the tools required to cost effectively characterize, target, and change patterns of energy demand are beginning to emerge. This dissertation consists of three chapters on the analysis of meter data and one on the policy implications of these new capabilities.

Chapter 2 analyzes hourly electricity and daily natural gas smart meter readings from 30,000 residential customers of Pacific Gas and Electric (PG&E). Meter data is used to derive distributions of previously unobserved characteristics of the housing stock. We show that the targeting of EE and DR programs would be improved through selection of households based on their positions within these distributions. It follows that every utility (or public utility commission) with sufficient metering infrastructure could apply similar techniques to improve the targeting, implementation, and evaluation of their energy efficiency and demand response programs.

Chapter 3 uses regression models of patterns in daily household electricity consumption to estimate the physical and operational characteristics of homes. We apply *semi-physical* regressors designed to capture patterns in the space-heating and cooling, scheduling, and occupancy of homes. When applied to data from approximately 160,000 PG&E customers, this approach supports an evaluation of competing regression model formulations and provides distributions of model coefficients used to evaluate patterns of domestic energy use, including annual and system peak coincident air conditioning loads, cooling set points, day of week scheduling, and lighting energy.

Chapter 4 presents a method for estimating the hourly timing of occupant driven loads based on smart meter data. The residuals of a predictive regression model are assumed to include *occupant activities* because occupant controlled energy use is not fully determined by externally observable factors. Occupant activity timing is converted into empirical distributions of the probability of such events by hour-of-day or day-of-week. With estimates calculated for approximately 25,000 PG&E customers and grouped using K-means clustering, prevailing patterns are interpreted as the result of occupant lifestyles — with applications in efficiency and demand response program targeting.

Drawing upon the applications developed in the preceding chapters, Chapter 5 discusses the potential for using smart meter data to support public interest utility programs in the context of ongoing concerns over public disclosure and privacy concerns, including malicious use by bad actors and inappropriate commercial use. We propose differentiated levels of access to meter data, with access to data mediated by *delegated analysis*, which allows stakeholders to receive the outputs of approved algorithms without requiring direct access to sensitive data. Such a system would provide privacy protection and oversight without foreclosing on creative and innovative uses of meter data.

Written with the memory of Professor Alex Farrell in mind. His wide-ranging interests, voracious reading, high standards, and intellectual honesty continue to guide and inspire my work. Dedicated to my newborn son, Owen Goggio Borgeson, whose gestation paralleled this work. His arrival has redoubled my commitment to leaving his generation a world in balance.

Contents

List of Figures

List of Tables

Acknowledgments

There is nothing quite so humbling as following your own ideas to their logical conclusions. I am grateful to have found the time, space, and, most importantly, gifted advisors and peers to support the revision of my ideas into this body of work. I owe a deep debt of gratitude to my advisor, quals committee member, and dissertation chair, Duncan Callaway, who supported every phase of this work. Despite a vague and changeable research plan, he had enough faith in this project to generously allocate time, thought, and funding to this work. I am also grateful to Ram Rajagopal at Stanford for welcoming me into his research group, sharing his exceptionally useful data, and, without any requirement to do so, generously allocating time to review my work, serve on my dissertation committee, and teach and advise me.

Until relatively late in this work, I lacked access to the quantity and quality of smart meter data I needed to perform the research as intended. I salute PG&E for sharing smart meter data under their control for research purposes and thank the people behind the Wharton Customer Analytics Initiative for compiling, sharing, and supporting a first-rate residential smart meter data set.

I owe thanks to Catherine Wolfram, for serving on my dissertation committee and meeting with me for hours to review and greatly improve several drafts of my work, and to Dan Kammen, for serving on my dissertation committee and helping to ground my research in urgent efficiency and climate policy work unfolding at the state and national level. I thank my qualifying committee members — chair Gene Rochlin, Duncan Callaway, Alan Meier, and Cris Benton — for their clarity of though on my proposed research and rejection of my initial scope of work. I also continue to be grateful for my mentors in Berkeley's Building Science program: Gail Brager, Ed Arens, and Cris Benton.

I have had the pleasure of collaborating with many talented students and researchers, including Brian Coffey, Omar Khan, Eric Kaltman, and the members of the EMAC lab at Berkeley; Paul Mathew and Phil Price at LBNL; and Adrian Albert, Jungsuk Kwac, and Amir Kavousian at Stanford.

I am indebted to the Switzer Foundation for funding my work and bringing me into their network of talented and passionate academics and professionals working on critical environmental problems. I am eternally grateful to Ren Orans, Snuller Price, and Jim Williams of E3 for funding the Alex Farrell Graduate Fellowship — which has deep personal meaning to me — and providing extremely valuable feedback on my work (especially on the most direct path to completion).

To my wife, Merrian Goggio Borgeson, who has supported me throughout this long process. She is a loving and supportive partner, top-notch energy analyst, gifted editor, and profoundly compassionate human being. To my mom, Barbara Lindsey, for a thorough review of my work and unwavering support of this and all preceding endeavors. I thank her for living her life in ways that manifest her compassion for people and love of the natural world. To my father, Gus Borgeson, for nurturing my curiosity, giving me the intellectual confidence to pursue my ideas, and teaching me not to take myself too

seriously. To my big sister, Hannah Borgeson, for providing professional caliber copy edits for the full draft of this work. She is a better urban environmentalists than I and taught me by example how to live well doing things I love.

This work would be greatly diminished without the support provided by any one of these people. They deserve much of the credit for what is good in this work and none of the blame for what is bad.

Finally, it has been a singular honor and pleasure to be a member of the ERG community. I am humbled by the breadth of interests, passion, intelligence, and generosity of the students, staff, and faculty. At ERG, I have been challenged to think about and practice interdisciplinary research, been introduced to profoundly important work dramatically different from my own, formed many lifelong friendships and working partnerships, and met the love of my life (now wife), the ERGie formerly known as Merrian Fuller. As I prepare to sever my formal ties to the program, I affirm that I am an ERGie for life.

Chapter 1

Introduction and summary

The infrastructure that delivers energy to utility customers is breathtaking in its scale and complexity, yet its purpose can be succinctly stated — ensuring that the supply of energy matches the demand for energy everywhere and at all times. Traditionally, utilities have seen their role as building and managing *supply*-side infrastructure to meet the exogenous demands of a growing and increasingly consumptive population. However, the energy crises of the 1970s raised concerns about energy availability and costs, and generated political support for efforts to reduce and better manage energy *demand*. Those efforts have taken the form of modern energy efficiency and demand response programs, which reduce overall consumption and shift the timing of consumption, respectively. True to their original motivation, the goals of these programs are primarily reducing costs and preventing grid failures. [1]

However, the utility industry is now facing the challenges of adapting its infrastructure to the needs of the 21st century. Infrastructure is aging, large power plants are difficult to finance and site, air-quality standards are increasingly stringent, and an efficient, clean, and renewable grid is a prerequisite for serious efforts to mitigate climate change.

The central question motivating this work is how energy efficiency and demand response programs can be updated and adapted to facilitate the evolution of utility infrastructure in the 21st century. Today's programs are largely designed around cost-containment objectives formalized in the 1980s, but energy efficiency and demand response are capable of providing planners flexibility to mitigate system capacity constraints, complement intermittent renewable energy integration, reduce conventional and climate pollution, decommission older plants, and avoid or defer expensive system upgrades.

It is widely recognized that meeting these new requirements will require significantly more energy efficiency and demand response resources, but it will also be important to control when and where the savings are achieved. The recent deployment of smart meters across a significant fraction of homes in many utility territories presents new opportunities for innovation and improvement of best practices.[2] Through data gathered by smart meters and software designed to manage and analyze ever-larger data sets, the new tools required to cost-effectively characterize, target, and change patterns of energy demand are beginning to emerge.

[1] Specific claims of costs and benefits remain the subject of debate. See for example [6, 77, 99, 22, 64].

[2] Smart meters are digital interval meters that are capable of recording and transmitting revenue quality electricity usage data with an hourly (or finer) timescale and natural gas with a daily time scale, and typically have other capabilities enabled by two-way communication with central utilities. They have been positioned as critical components of smarter grids, which will make end-to-end use of information technology (in the form of measurement, communication, data processing, and controls) to improve the flexibility and efficiency of grid operations. In this context, smart meters provide detailed temporal and spatial information about grid conditions and patterns of energy demand and serve as gateways for customer communications.

1.1 Summary of this work

This work develops smart meter data analysis tools and explores their potential to increase the scale of energy efficiency and demand response programs while improving the precision of their targeting and the reliability of their outcomes. The work is presented in four main chapters, each of which focuses on a different application of meter data. The chapters are complementary but are also designed to be read as stand-alone works. For this reason, they each feature their own introduction and conclusions and duplicate background information where necessary.

1.1.1 Program targeting using meter data

The most important difference between smart meter data and the monthly meter readings they replace is their improved time resolution. With hourly interval readings, it is possible to resolve daily load shapes, peak and minimum demand, and differences across days of the week or months of the year. **Chapter 2** presents an empirical analysis of smart meter data from a random sample of 30,000 residential customers in the service territory of Pacific Gas and Electric (PG&E). The main goal of this chapter is to demonstrate the potential for this data to reveal distributions of previously unobserved characteristics of the housing stock, with an eye toward applications in the targeting of energy efficiency and demand response programs. The analysis relies on hourly readings of electricity demand spanning at least 6 months and an associated zip code for each residence. Derived metrics are used to distinguish among homes with electric and gas heating and those with and without cooling loads. We estimate base and variable loads and characterize intra-day and seasonal variability. We also identify prevailing patterns in the timing and intensity of household peak demand.

The savings and flexibility achieved by energy efficiency and demand response programs are contingent on finding and enrolling customers whose ability to participate is well matched to program objectives. Currently, program designers must choose between the higher costs associated with finding and enrolling customers in more tailored programs or the lower customer acquisition costs but also lower performance associated with more prescriptive, mass-marketed programs. The balance of incentives has historically skewed toward prescriptive measures, like more efficient light bulbs, but the growing interest in achieving deeper and more reliable savings is directing the attention of program managers toward more effective program targeting.

Based on the non-uniform distributions of observed characteristics, energy efficiency and demand response programs stand to benefit from using meter data to target program participants. This analysis shows that in the PG&E service territory, the energy used by the least consumptive 50% of homes is equal to that of the top 10%. In 45% of households, more than 40% of annual energy use can be attributed to loads sustained at overnight minimums. This number is less than 20% of annual energy in fewer than 5% of homes. This means that base loads comprise large fractions of annual residential

energy use for the vast majority of residential customers and suggests that better timing of controls and reduced standby losses from plug loads have significant potential to deliver energy savings. Typical peak hours of residential demand span from 6 to 8pm, and household peak demand tends to fall on either the warmest or coldest days. In addition to being more energy intensive, summer loads are also more variable. Along with these specific insights, we have also developed a set of definitions and distributions of metrics that should be useful for purposes of multidimensional benchmarking and comparison among residences. It follows that program administrators in every territory with sufficient metering infrastructure could apply these or similar techniques to improve the targeting, implementation, and evaluation of energy efficiency and demand response programs.

1.1.2 Semi-physical regression modeling

With improved temporal resolution of meter data comes the ability to more precisely model the factors that drive consumption, including weather, day length, and scheduling. **Chapter 3** is focused on the use of regression models to identify patterns in household electricity consumption and provide information about the physical and operational details of homes. These insights can better inform the planning, execution, and evaluation of energy efficiency and demand response programs. Models of this type are known as inverse models because they are designed primarily to estimate the operational parameters that determine consumption rather than to predict consumption given operational parameters. We develop a set of *semi-physical* regressors intended to capture patterns consistent with the space heating and cooling, scheduling, and occupancy of homes. These regressors are combined in regression models to test the explanatory power of individual regressors as well as to capture, via variations in model structures, the anticipated range of operational strategies found in homes. The models are fit using smart meter data drawn from a stratified sample of 160,000 homes in PG&E's service territory. This approach provides rich distributions of modeled outcomes that can be used to evaluate patterns of domestic energy use, including annual and peak coincident air conditioning loads, cooling set points, day-of-week scheduling, and lighting loads.

Residential characteristics inferred from model fits are used to estimate patterns in consumption by day of week, the magnitude and distribution of heating and cooling loads, and distributions of heating and cooling set points. Household characteristics are then used in several example analyses that demonstrate tools to target heating and cooling efficiency programs, estimate potential savings from set point changes, and identify differentiated household contributions to peak grid demand. The top 10% of residences in the sample were responsible for 39% of cumulative annual cooling energy, and the top 25% were responsible for 70%. By contrast, the lowest 50% of households contribute just 6% of total annual cooling demand. Across-the-board increases in cooling set points of 1, 2, and 5°F are found to decrease annual cooling loads in the sampled residences by 10, 19, and 44%, respectively. Roughly half of the sample's daily household energy demand coincident with peak days of grid demand comes from just 10% of the sample. All of these statistics

strongly suggest the benefits of targeting programs to reach the households whose program participation would most benefit demand-reduction objectives.

1.1.3 Estimating the timing of occupant-driven loads

Along with the increasing availability of smart meter data, one of the most exciting recent developments in energy efficiency programs is that they have successfully implemented programs that take into account insights from the behavioral sciences. By capturing whole-home energy consumption, which includes loads directly controlled by occupants as well as automated and scheduled loads, meter data can provide ground truth data to support behavioral research. Lessons from this rapidly developing field of research stand to significantly improve behavioral programs. An important early lesson from this work is that even after controlling for all known externally observable influences, patterns of consumption are extremely diverse. This is consistent with the understanding that household composition, individual preferences, and appliance ownership are similarly diverse and implies that programs designed to embrace this diversity will outperform those that ignore it.

Chapter 4 explores methods for estimating and classifying patterns in the timing of occupant-driven loads using smart meter data. We develop a regression model designed to control for loads that are predictable in time and outside temperature. This model has separate hourly terms for piecewise temperature responses and hour-of-week fixed effects. The resulting residuals are assumed to be composed of a combination of typical model errors and isolated "*occupant activities*" poorly fit by the predictive model because their timing and magnitude are determined by the needs of unobserved occupants.

On the assumption that large positive model residuals will tend to be caused by occupants operating appliances and electronics, occupant activity timing is converted into empirical distributions of the probability of such events by hour of day or day of week. With such estimates calculated for thousands of homes and then grouped using K-means clustering, several prevailing patterns of occupant activity emerge and are interpreted as lifestyle choices through example applications in efficiency and demand response targeting. For example, the majority of households demonstrate increased activity on the weekends, but some demonstrate decreased activity. This classification very likely separates households that enjoy time at home on weekends and those that tend to spend weekends away. Clusters of activity by hour of day similarly reveal several typical patterns reflecting the tendency of the majority of households to be most active in the early evening, but with others peaking mid-morning, mid-day, and overnight. Prevailing patterns in occupant activity are then compared to hour-of-day cooling demand to find mismatched schedules. The expected magnitude and probability of occupant-driven activities are used to identify members of the sampled population most and least active during expected grid peak demand periods.

1.1.4 Balancing beneficial uses with privacy concerns

The details of consumption that support the beneficial uses of meter data come hand in hand with the potential for the data to reveal private information about individuals or households. The magnitude of consumption, patterns of occupancy, presence or absence of specific loads, and any number of context-sensitive details of consumption can all be seen in meter data. Such details could be used by law enforcement or national security agencies to develop profiles (accurate or not) of suspicious activity, by bad actors to learn details about specific households, or by commercial entities to support unwanted profiling and marketing. Unlike online shopping accounts, web-based email accounts, or social media site membership, all of which can be also used for the above purposes, smart metering is virtually mandatory and ostensibly dedicated to achieving public interest goals. While realizing the many beneficial uses of meter data, regulatory agencies are also obligated to take responsibility for protecting the rights and privacy of utility customers.

Chapter 5 discusses potential public benefits from the analysis of smart meter data in the context of ongoing concerns over public disclosure and privacy concerns raised by the data. The cost-benefit analysis of smart meter deployment is typically focused on benefits from automated meter reading and developing the infrastructure for real-time pricing. However, these benefits could be exceeded by savings from better understanding the patterns of demand relevant to energy efficiency and demand response programs. Achieving these additional public benefits will require a process of ongoing learning and innovation in the field of meter data analysis and careful consideration of rules governing data access and customer privacy. In particular, the trend toward private companies serving as custodians of personal data and recent examples of apparently anonymized data being re-identified through correlations with public data sets should be addressed by regulators.

This work proposes differentiated levels of access to meter data, with some access mediated by *delegated analysis*, which would allow third parties to specify algorithms for data analysis and receive their outputs without being granted direct access to sensitive data. Such a system would provide ample room for oversight without foreclosing on creative and innovative uses of meter data. We draw upon our experience working with a representative sample of meter data from a major utility's service territory to provide examples of analyses that could be delegated to a trusted party and returned without personally identifying information or loss in the value of the end results.

1.2 Data sources

The data sets of smart meter readings and limited customer account information that supported the analyses documented in this dissertation both originated from residential accounts in the service territory of PG&E. One was provided to our research partners in Professor Ram Rajagopol's Stanford Sustainable Systems Lab directly by PG&E. The other was developed by the Wharton Business School's Customer Analytics Initiative

and made available through an open call for research proposals. This section provides background information on these data sets as well as the supporting weather and demographic data collected to complement them. Specific chapters go into additional details as necessary.

1.2.1 Stanford collaboration data

The raw sample consisted of meter readings from approximately 180,000 residences across 588 five-digit zip codes. The sample was stratified to ensure well-specified numbers of accounts associated with single-family residences, landlords, and renters. Basic information including residence zip code, service type, and rate plan were provided for each account. Validation rules mandating at least 180 days worth of data, no more than 15% of readings equal to zero, and mean power demand greater than 110W narrowed the results to approximately 160,000. For applications requiring estimates of annual consumption, the sample is further reduced to 128,000 households. The data covers April 2008 to the end of November 2011.

According to PG&E's 2009 SEC 10K filing, [90], the company served nearly 4.5M residential customers, or about 28 times the number of households in our sample. In 2009, all of PG&E's residential customers consumed approximated 31,000 GWh of electricity, which is 36% of PG&E's total for the year. This is an average of nearly 6,900 kWh per household. The average for the validated data sample used for this study is 6,960 kWh/year — slightly higher, most likely due to the validation rule that eliminates extremely low usage.

1.2.2 Wharton Customer Analytics Initiative data

This data was provided via an open call for research proposals by the Wharton Business School's Customer Analytics Initiative.[3] The data set consists of records for 30,000 PG&E residential customers, with 10,000 randomly sampled from each of three geographic zones — Coastal, Inland Hills, Central Valley — that roughly cover the territory's climate variability. The populations of the zones are: Coast: 1.1M, Central Valley: 1.5M, and Inland Hills: 1.8M. For every sampled account, electricity and natural gas smart meter data was provided along with account information that included zip code, service type, and rate plan; anonymized billing and payment data; past customer service interactions; and limited data on utility program participation.

Because they were randomly sampled, the meter counts correlate with population densities, so the highest counts are found in the San Francisco Bay Area,and around other population centers along the coast and in the Central Valley. Of the original 30,000 customers, roughly 24,700 meters passed all data validation and cleansing criteria. Validation required more than 180 days of observations, no protracted periods of zero energy

[3] `http://www.wharton.upenn.edu/wcai/`

consumption, and average power demand > 110W.

1.2.3 Weather

To support the study of space conditioning and the impact of housing stock and household characteristics on energy demand, weather data was gathered for each of the zip codes in both data sets. Hourly temperature, pressure, and humidity observations spanning the period of available readings for each zip code were collected from the online weather data aggregator web site Weather Underground according to the following process:

1. Each zip code was sent to the Weather Underground API to return a list of closest non-airport stations (distances are reported in km). For example:
   `http://api.wunderground.com/auto/wui/geo/GeoLookupXML/index.xml?query=94611.`

2. Because data formats are varied and airport stations are few and far between, non-airport station data was parsed, with attention to the reported distance from the submitted zip code (distances in km are provided by the API).

3. Nearby stations were polled using simple metrics of data integrity. The three closest stations that passed integrity checks were selected for data download.

4. Interval data on Weather Underground is available only one day at a time, so a threaded crawler was used to assemble a full data set per station.

5. Due to bad data associated with noisy data feeds and intermittent sensor failures or misreads, extreme values, for example temperatures less than -20°F or greater than 120°F (aa range that CA weather falls within), were excluded from the recorded data.

6. Separate data files were saved for each station's raw data. The raw interval data was post-processed into hourly averages for each station.

7. The hourly data for each of the three stations closest to the center of each zip code were then averaged with missing values excluded to give the best estimate of temperature time series for each location.

8. The zip code and time stamps associated with each household's meter readings are used to match weather data to household consumption as needed.

1.2.4 Census data

Zip code aggregated demographic data was extracted from tables DP02, DP03, DP04, and DP05 of the Zip Code Tabulation Area (ZCTA) summaries of the American Community Survey 2009 5 year extract, [114]. ZCTAs are not the same geographic regions as zip codes and the Census Bureau no longer provides an official mapping between the two.

For this reason, ZCTAs were mapped to zip codes using a well-respected, but unofficial, crosswalk file from 2011 available from:
`http://udsmapper.org/ziptozctacrosswalk.cfm`.
Census data was then matched to household zip codes on an as-needed basis during the course of data analysis.

Chapter 2

Empirical targeting of residential efficiency and demand response programs

2.1 Introduction

This chapter presents an analysis of several applications of electricity and natural gas smart meter data from 30,000 anonymous residential customers in PG&E's service territory. This work is structured as a survey of potential program targeting applications of smart meter data for three main reasons. First, such data has only recently been made available for research purposes. This early analysis of smart meter data is necessarily exploratory in nature. The metrics we use establish the rough contours of problems that can be addressed using smart meter data and offer guidance for future, more technical, work. Second, the data set consists of hourly meter readings, with only a limited amount of information embedded in them regardless of analytical methods. Third, the highest use of the fruits of such analysis is in their application in practice and at scale. Metrics that are easy to communicate, implement, and compute will be critical to scaling up programs that make use of empirical targeting methods.

We present findings on territory-wide distributions of electricity and natural gas consumption intensities; correlations among usage, household features, and demographics; patterns in location and time of heating and cooling demand; the timing of household level peak electricity demand; the contribution of base load to annual consumption; distributions of daily load shape characteristics; and spatial patterns of demand during periods of grid stress. Our results strongly suggest that information derived from smart meter data at very low cost can be used to direct and monitor grid management goals, and to better plan, implement, and evaluate efficiency and demand response programs. While simple calculations are capable of providing some of these benefits, we conclude that more sophisticated and specialized analytical tools should be developed to capture additional value from meter data.

2.2 Background

Smart meter electricity readings are typically recorded at 15-minute or hourly intervals. This is not fast by the standards of modern sensing and logging equipment, but it is roughly three orders of magnitude faster than the monthly readings that have traditionally supported utility billing systems. It is also well matched to the task of characterizing home energy consumption that is subject to diurnal patterns in usage, temperature, and occupancy. While additional applications can be devised using data measured at minute, second, or even kilohertz time scales [41, 62, 35], the essential patterns of home energy are visible when there is sufficient data available to resolve daily load shapes. In contrast, natural gas readings are typically recorded only once per day. Natural gas is consumed by fewer end uses in a home (primarily hot water, gas dryers, cooking, and space heat), and daily data is sufficient to observe variations due to fluctuation in occupancy and weather

and to estimate typical base usage [1].

As of May 2012, it is estimated that electric utilities in the US had deployed 36 million smart meters (covering 1 in 3 households), with 65 million projected to be deployed by 2015 [46]. As of this writing, as many as two dozen service territories in the US have effectively achieved full coverage in their deployments[2], with densities sufficient to support the analysis of patterns of energy demand in unprecedented temporal and spatial detail.

The related problems of predicting and explaining commercial and household energy usage are notoriously tricky. Buildings are engineered systems but typically suffer from defects in construction and controls. While they are operated within specific comfort and environmental quality bounds, they are ultimately at the service of people with hard to predict occupancy, needs, and desires. They also contain a wide variety of electronic appliances and equipment whose usage has little or no correlation to the purposes of shelter and comfort that define traditional building functions. Even when analysis is restricted to heating and cooling loads, i.e., the response of engineered systems to deterministic thermal processes, the timing and magnitude of building energy consumption are difficult to predict [84, 113, 9, 16]. This is because buildings store and exchange thermal energy in many forms across multiple time scales. These can depend on building geometry, furnishing details, the massing and thermal conductance of construction materials, current and past weather conditions, the temperature of the ground beneath the building, and even surrounding vegetation. Due to their complexity, many of the thermal properties of buildings are estimated rather than measured. For example, buildings allow infiltration of outside air via unseen cracks, gaps, and spaces in construction — all of which sum to a total effective infiltration rate, which is also conditional on wind speed and direction.

Work using smart meter data necessarily turns the above formulation on its head. There is a reliable and growing supply of whole home electricity usage data observed at roughly 15-60 minute intervals (and often daily gas usage and potentially water usage), so what questions can such data address? What processes can tolerate the uncertainties inherent in drawing conclusions using coarse and anonymous data? One answer is that smart meter data can improve the targeting and evaluation of efficiency and demand response programs. This is partially because these programs have not traditionally incorporated any detailed consumption data at all — demand response already has moved more in this direction than energy efficiency. The programs that have tend to be constrained in size by the overhead of procuring and analyzing that data. The other reason smart meter data is a good fit for energy efficiency and demand response programs is that such programs are probabilistic and fault tolerant in nature — even imprecise information can be used to improve program targeting and execution, lower costs, increase the magnitude of impacts, and reduce aggregate uncertainties. This is because both energy efficiency and

[1] In general, the faster the sampling rate, the better the usage can be disaggregated. Fast sampling comes with additional data logging, transmission, and storage concerns and is far more problematic from a privacy perspective.

[2] In the face of backlash against smart meter installation some utilities have delayed installations or instituted an opt-out program that precludes absolutely full coverage.

demand response seek aggregated outcomes from individual actions. As long as a margin of error is built in, not every project is required to go as planned to achieve aggregate kWh or kW reduction goals.

2.3 Prior work

Prior work relevant to this study can be assigned into several different categories: (1) large surveys that seek to categorize energy consumption by end-use, (2) studies of methods for disaggregating thermal loads, scheduled loads, and occupant controlled loads from whole building data, (3) studies involving interval data and sub-metered end uses, (4) analyses of large samples of utility customer meter data. The fourth category of studies has only recently become viable due to widespread smart meter deployment and the development of research tools adapted to large data set.

The most prominent characterization of residential energy for the US is the EIA's Residential Energy Consumption Survey (RECS), which captures a great deal of information about various residential end uses, fuel sources, and occupant and physical characteristics [31]. However, RECS, which is updated every 4 years, is based on a sample of about 12,000 homes. It is focused on home and household characteristics, with annual energy use broken out into end uses, but lacks the spatial and temporal resolution required to support more localized and application-driven studies. In California, efficiency program initiatives are supported by detailed studies of appliance ownership and operation based on sales data and extensive, labor intensive surveys. For example, the California Residential Appliance Saturation Survey, [51], is based on a survey of approximately 25,000 CA utility customers. For many avenues of inquiry, the improved resolution and widespread availability of utility meter data dramatically improves the ability to characterize patterns of energy use.

It is not news to building researchers that meter data can speak volumes about energy use in individual and aggregated buildings. Dating all the way back to the seminal study of home energy in the Twin Rivers housing development outside Princeton in the late 1970s, [108] and the subsequent development of the PRISM method of heating and cooling regression analysis, as described in [33] with subsequent extensions by [57, 40, 52, 96], utility data has been recognized as a key to understanding the dynamics of energy use in practice. For just as long, it has also been understood that the preferences and particular behaviors of occupants are often more important to determining energy consumption than predictable thermal processes, as documented in [109, 68, 48]. In sum, both deterministic and non-deterministic processes contribute to consumption and meter data should be modeled and interpreted accordingly.

Because of the value of ground truth data, utilities and utility commissions have made significant investments in data gathered via interval metering as part of their grid planning and management processes as well as in the course of planning and evaluating efficiency programs. These studies often include separate metering of end uses or appliances and

surveys or other contact with occupants. Through their work with interval meter data from hundreds of households, they anticipate many potential applications of smart meter data. For example, [88] describes the ELCAP program in the Pacific Northwest, which gathered sub-metered interval data along with physical, occupant, and appliance characteristics for 454 residences and 140 commercial buildings between the late 1980s and early 1990s to improve efficiency program planning and improve load forecasting models. [87] documents a 2003 study that recorded sub-metered interval data for 204 residences in Central Florida to improve efficiency program evaluations and develop empirical appliance load profiles. [107] describes a 2002 study using sub-metered interval data from 400 residences across Denmark, Italy, Portugal, and Greece to assess appliance efficiency potential. This study documented a significant difference in appliance load characteristics between countries, but found universally high energy use by the standby power consumption of electronics. Finally, [11] disaggregates smart meter data from 327 households in Ontario, Canada into base, heating, cooling, and active load categories and suggests its utility in targeting utility efficiency and demand response programs.

Recently, larger data sets have been analyzed as well. Municipal interest in climate mitigation has resulted in collaborations with researchers to study utility billing data. [55] analyzes monthly electricity and gas data from 6,500 customers of all types in Cambridge, MA and evaluates predictive models of consumption based on building level tax assessor and GIS data. [45] uses data from newly mandatory energy disclosure requirements to study 10,000 multi-family and commercial buildings in New York City to predict energy consumption by fuel type and cluster buildings with similar consumption characteristics. Efficiency programs can also accumulate sufficient information to publish studies of whole populations of buildings. For example, [97] use the outcomes of nearly 5,000 energy audits in Texas homes to discuss distributions of HVAC equipment sizing and operational performance relevant to assessing efficiency program potential. [10, 2, 60] use detailed meter data to develop classes of customers based on similar patterns of usage.

However, at the time of this writing, there are no studies of both large and representative sample sizes and sub-daily meter readings that characterize the patterns of usage across a service territory. Because smart meter interval data is capable of supporting loads shape analysis and disaggregation into load categories and smart meters can be sampled to be representative of an entire service territory, such studies can contribute valuable insights into the nature of electricity and natural gas demand and inform the design, targeting, execution, and evaluation of efficiency and demand response programs. This work is a demonstration of several methods of analysis of smart meter data that provide insights relevant to the improvement of demand-side-program outcomes.

2.4 The data set

The data set used for this study come from Pacific Gas and Electric (PG&E), which serves more than 5 million customers, including 4.5 million residential accounts, in a

service territory spanning most of northern and central California. The data consists of records for 30,000 residential customers, with 10,000 randomly sampled from each of three geographic zones — Coastal, Inland Hills, and Central Valley. The coastal climate is mild, the Pacific Ocean moderating temperatures and enforcing a winter rainy season, with long, sunny, and warm summers and significant fog coverage. It includes the cities Eureka, San Francisco, and Santa Cruz, and nearly reaches Santa Barbara. The inland hills see larger diurnal and seasonal temperature swings, but only the eastern segment of the hills experiences freezing conditions and snow routinely during winter. This region includes Napa and Sonoma Counties, San Jose and the Silicon Valley, the East Bay hills, the Santa Cruz mountains, and, farther east, parts of the foothills of the Sierra Nevada range. The Central Valley, on the other hand, is one of the most productive agricultural areas on earth and features the required climate conditions: long, sunny, and hot summers and mild winters. It includes the cities of Chico, Sacramento, Fresno, and Bakersfield. Once customers were selected, any smart meter electricity and natural gas data associated with their account, along with their billing data, zip code, and a subset of records of customer contact with the utility via rebate programs, phone support, etc. were gathered. This data was provided for study via an open call for research proposals by the Wharton Business School's Customer Analytics Initiative[3].

To support the study of space conditioning and the impact of housing stock and household characteristics on energy demand, additional data was gathered for each of the 823 zip codes in the data set. Hourly temperature data for each zip code was collected from the online weather data aggregator web site Weather Underground, with simultaneous values from 3-5 weather stations averaged, dropping missing observations, to produce an aggregate weather data time series for every zip code. Zip code aggregated demographic data was extracted from Zip Code Tabulation Area (ZCTA) summaries of the American Community Survey 2009 5-year extract, [114].

[3] `http://www.wharton.upenn.edu/wcai/`

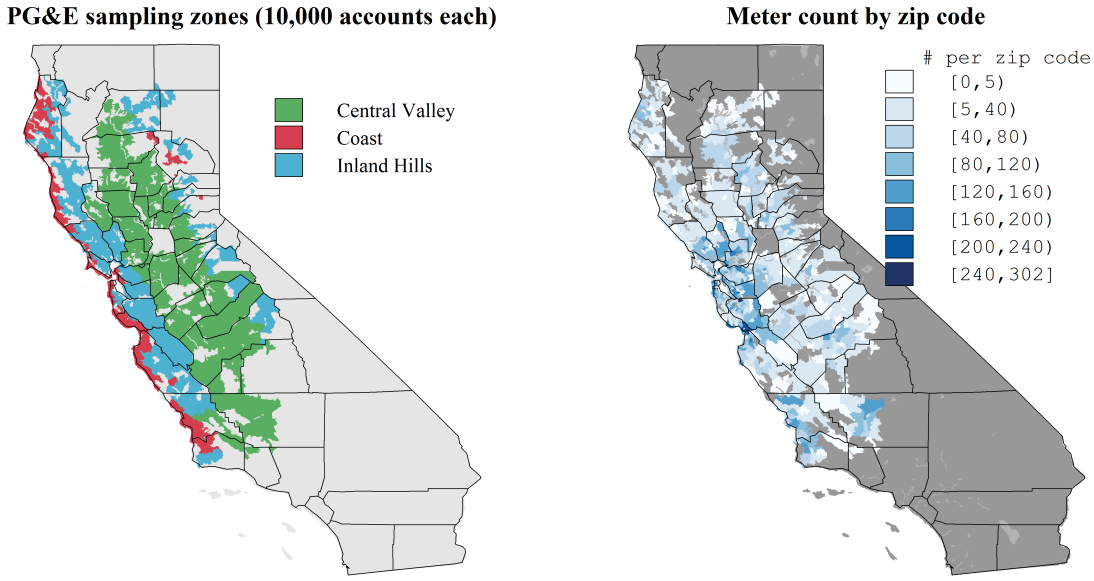


Figure 2.1: Maps of the three account sampling areas, from which 30,000 accounts were randomly selected - 10,000 from each zone (left) and meter counts by zip code (right).

Figure 2.1 on page 16 presents maps of the PG&E service territory covered by the data set, depicting the Coastal/Inland Hills/Central Valley geographic zones used to partition the sample of 30,000 customers into 10,000 from each zone (left) and the count of sampled smart meters per zip code (right). The populations of the zones are: Coast: 1.1M, Central Valley: 1.5M, and Inland Hills: 1.8M. Because they were randomly sampled, the meter counts correlate with population densities, so the highest counts are found in the San Francisco Bay Area and around other population centers along the coast and in the Central Valley. Of the original 30,000 customers, roughly 24,700 meters passed all data validation and cleansing criteria. Validation required more than 180 days of observations, no protracted periods of zero energy consumption, and, to eliminate samples from unoccupied homes or faulty meters, average power demand $>$ 110W. Further validation is required for certain applications, such as calculations requiring a full year of data. Approximately 22,300 electricity customers and 16,000 natural gas customers have data spanning a year or more. Also, some meter data is discarded when matching meter data to rarely incomplete temperature observations from nearby weather stations. Where figures or discussion rely on data from a number of residences significantly different from the usable sample of 24,700, the actual numbers will be noted.

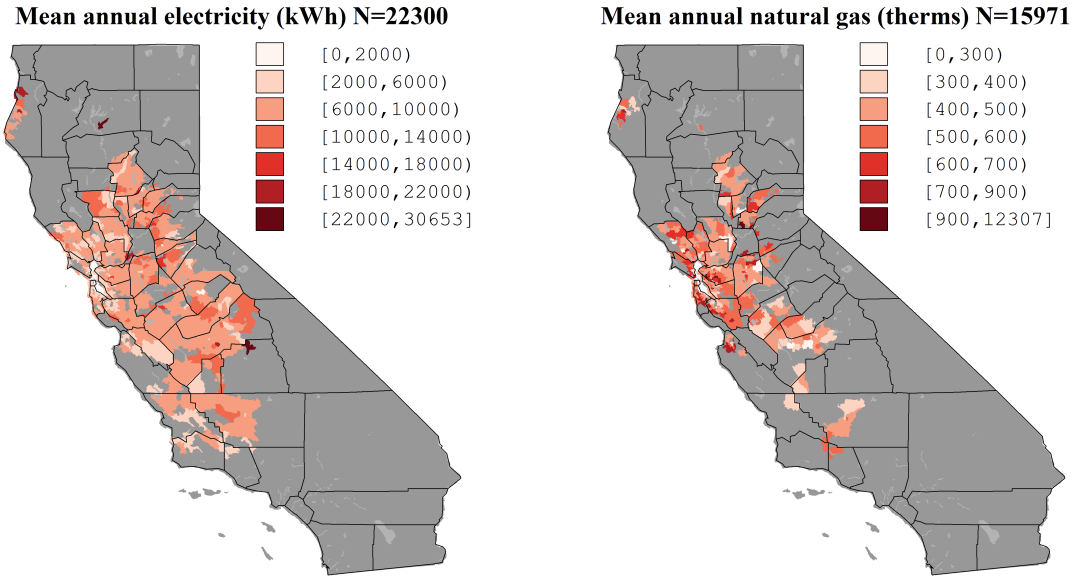


Figure 2.2: Maps of the mean annual electricity consumption in kWh per zip code (left), and the mean annual natural gas consumption in therms per zip code (right).

Figure 2.2 on page 17 presents maps of mean annual electricity (left) and natural gas (right) consumption by zip code. The natural gas data covers fewer zip codes than the electricity data, reflecting both instances of electricity-only PG&E customers and the more modest roll-out of natural gas smart metering. Temperatures follow a general pattern of cooler coastal weather, warmer weather in the large Central Valley, and cooler conditions again in the Inland Hills. The consumption maps reveal climate-driven patterns of electricity and natural gas usage consistent with higher cooling needs in the Central Valley and higher heating needs in the Inland Hills. However, these patterns do not consistently dominate over other sources of variability.

2.5 Analysis

What factors help predict patterns of residential energy consumption? How can these factors be used to improve our understanding of which buildings are good candidates for energy efficiency interventions or well suited for enrollment in demand response? In this section, we present several analyses, each focused on a different aspect of consumption. These analyses are intended to underscore the breadth of useful targeting information that can be extracted from meter data.

2.5.1 Distributions of annual energy use

In the broadest terms, the goal of this research is to better characterize the stock of residential buildings inorder to understand the patterns of consumption that are indicators of high or low energy use and help understand what causes the differences. The most basic tools in this work are histograms, density plots, and cumulative distributions that illustrate distributions of energy use or other features of consumption. These distributions highlight the diversity of usage and, especially when broken down by specific geographic, demographic, and housing stock characteristics, invite benchmarked comparisons between households. Residential attributes capable of predicting placement within such distributions are useful in program targeting and can also be used to provide information and feedback to residents.

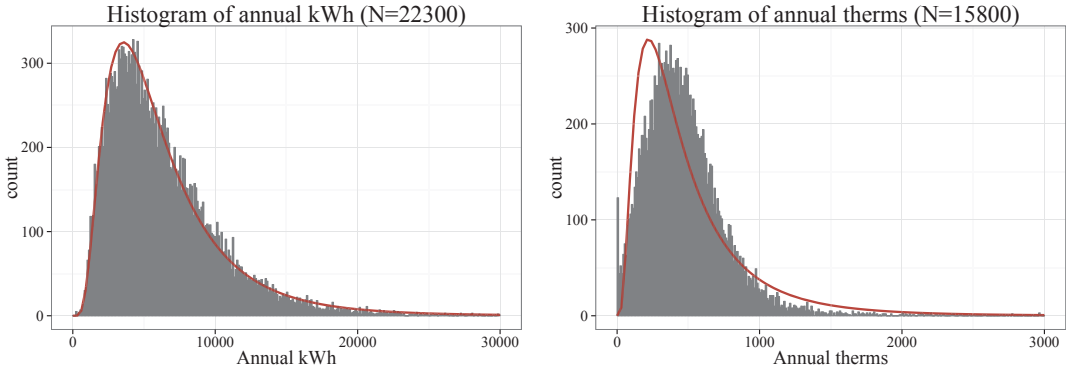


Figure 2.3: Distribution of residential electricity use (left) and natural gas (right) across the entire sample of homes. The samples are naturally anchored on one side at zero (net PV production is not included) and feature fat tails skewed toward higher energy use. As seen through the red line log-normal fit, the electricity distribution is roughly, but not exactly, log-normal (geometric mean $\bar{x}^*$=5300 kW and multiplicative standard deviation s^*=1.88 for comparison to others in [65]). The natural gas distribution is not as well fit by a log normal distribution. The fit shown has $\bar{x}^*$=386 therms, s^*=2.12.

Figure 2.3 on page 18 provides a histogram and log-normal fits for the annual energy use of the approximately 22,000 meters with at least a year's worth of electricity data (almost 16,000 for natural gas). The distributions are anchored at zero and skewed with fat tails toward higher consumption. This shape is consistent with the skewed distribution first documented in 1987 in commercial premises [13], and found in the data from Cambridge, MA [55]. As the log-normal fit lines demonstrate, the distributions are roughly, but not exactly, log-normal, with a better log-normal fit for annual electricity. It is possible to interpret the form of statistical distributions in terms of the properties of random variables that give rise to them. In this context, it is interesting to note that log-normal distributions often arise as the product of independent, normally distributed variables

and that energy use is the product of the level of service demanded and efficiency of the equipment that provides the service. This way of thinking about consumption is different from the more conventional view that total consumption is best understood as the sum of various end uses.

The mean values in the data (6,457 kWh/year and 484 therms/year) are consistent with the EIA estimates (6,888 kWh/year and 400 therms/yr) for the whole state in 2009 [31] and with PG&E's mean residential sales (6,953 kWh/year and 482 therms/year for 2009, according to [90]). PG&E's service territory is in northern California, where there are higher heating loads, so it is unsurprising that the average for natural gas is somewhat higher than the state average.

2.5.1.1 Cumulative distributions of annual consumption

Cumulative distribution of annual electricity (left) and natural gas (right) usage

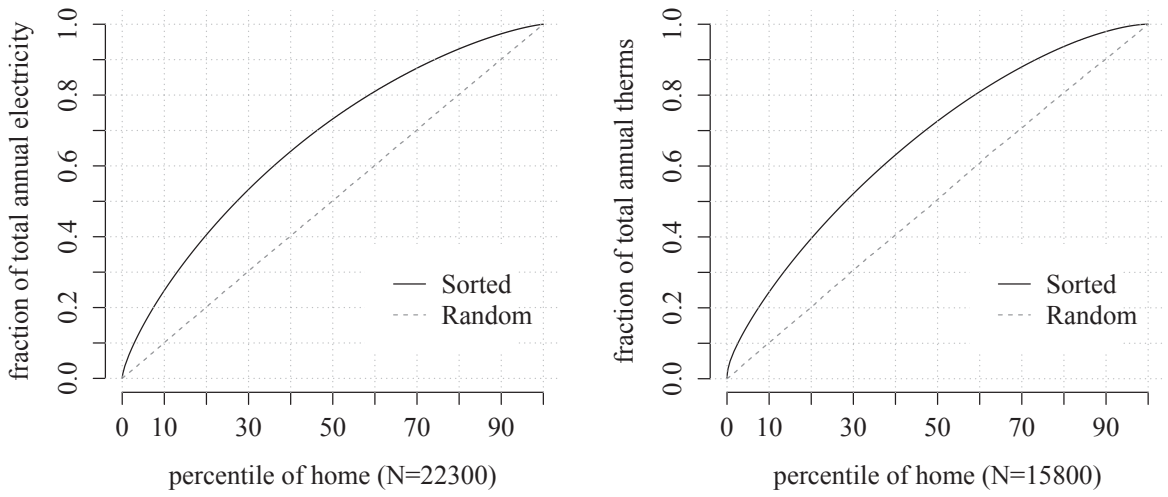


Figure 2.4: Cumulative annual electricity (left) and natural gas (right) usage across the entire sample, sorted from highest energy users to lowest.

The same underlying data can be sorted from greatest to least annual energy use and presented as a cumulative sum of all the energy consumed by the sample. In Figure 2.4 on page 19, the cumulative sum is plotted for annual kWh and therms, with the straight line cumulative sum that would result if the households were instead selected at random from the population. Here it can be seen that the top (leftmost) 10% of homes use approximately 25% of the total electricity. It can also be verified that the lowest (rightmost) 50% of homes use just over 25% of total electricity. Similar values apply to natural gas usage. To provide context to these percentages, PG&E's total residential

sales for 2009 were 31,200 GWh and 1.95 Billion therms across 4,492,359 and 4,046,364 customers respectively. The observation of uneven distribution of consumption across residential consumers strongly suggests that targeted selection of homes for participation in energy efficiency programs should be capable of handily out-performing those based on random or self-selection. The mean consumption of the top 10% is 2.5x the sample average, so targeting a program that achieves savings proportional to total consumption would be expected to yield 2.5x the savings of a program applied to a random sample of 10% of the population.

2.5.1.2 Co-variation between electricity and natural gas consumption

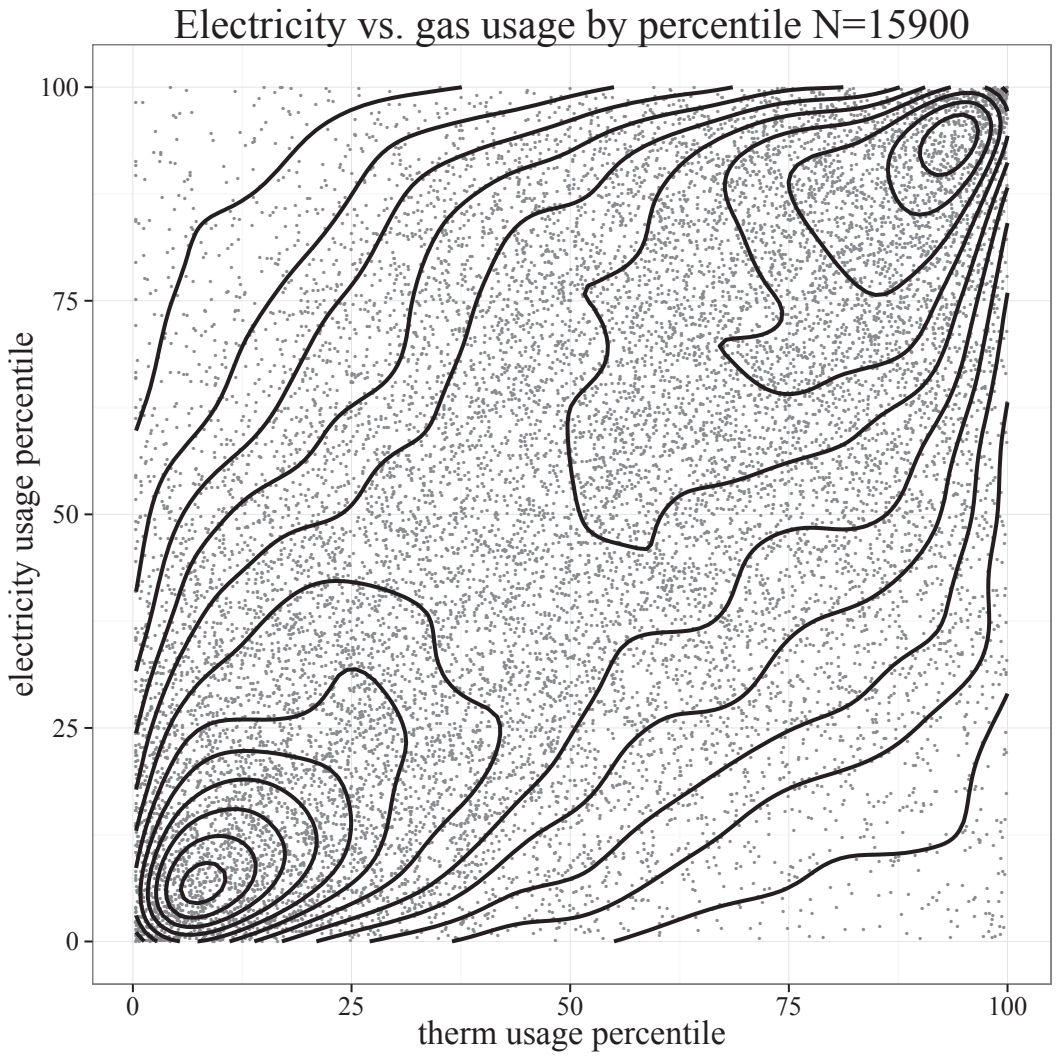


Figure 2.5: Household percentile of annual kWh vs. percentile of annual therms.

The observation of qualitatively similar distributions of annual kWh and therm consumption raises the possibility that the same homes are high consumers of both types of energy. It is not hard to imagine physical characteristics (like square footage) or household characteristics (like the number of occupants) that would tend to increase or decrease both types of usage. Figure 2.5 on page 20 is a scatter plot, with each point representing a household. The x-axis is the household's percentile in annual natural gas consumption. The y-aixs is the household's percentile in annual electricity consumption. Contour lines of equal point density have been added to help interpret the pattern of points. The closer a point is to the x=y diagonal line, the closer its percentiles of electricity and natural gas are.

The plot reveals that there is a general correlation between the electricity consumption and natural gas consumption (the simple correlation between percentiles is 0.52 and the simple correlation between values is 0.39), but the high degree of scatter also indicates that there are many other factors that shape household consumption. An obvious example is that some households have electric hot water heaters, skewing a household's usage toward a higher percentile of electricity usage and a lower percentile of natural gas usage. In this way, a home's location in the plot reveals information that might be used to better plan efficiency programs with different options for electricity and natural gas savings.

2.5.2 Household characteristics

In this section, we are concerned with how well externally observable characteristics predict (or at least correlate with) annual energy use. Any characteristic with predictive power can be used to better design and target programs. For example, conventional wisdom holds that energy use increases with household income and physics-based models of consumption are frequently based on the assumption that much of a household's energy will be put toward heating and cooling. However, there can still be a great deal of variation in residential energy use even after controlling for these factors.

Because our data is restricted to meter readings and zip code for each residence, we do not have sufficient information to assign physical or demographic information individually. However, through the American Community Survey, we have several household features aggregated for each zip code tabulation area. Here we present the results of comparing zip code averaged annual kWh and therms against zip code aggregated census data.

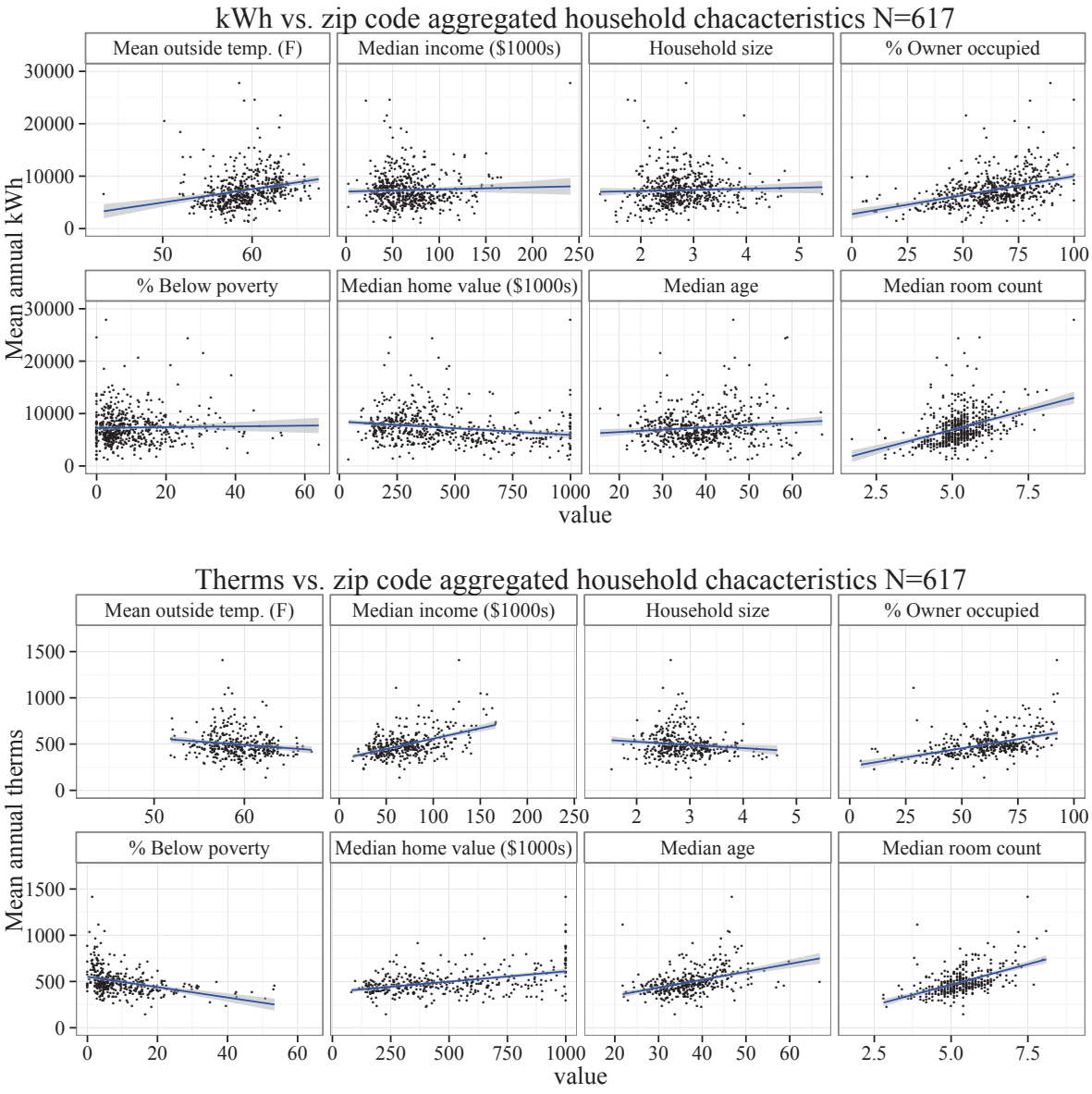


Figure 2.6: Zip code average annual electricity and natural gas usage vs. relevant zip code level demographic indicators.

Figure 2.6 on page 22 presents scatter plots of zip code averaged consumption of electricity (top) and natural gas (bottom) against outside temperature and a selection of 7 household characteristics, also aggregated across each zip code. Every plot includes a regression line with its 95% confidence interval. Outside temperature appears to have a significant correlation with both natural gas and electricity, as do the percentage of homes that are owner occupied and the number of rooms in homes. However, median

population age and financial indicators, including the percentage of households below the poverty level, median income, and median home value, are more strongly correlated with natural gas usage than electricity. Household size does not correlate well with either electricity or natural gas consumption.

Annual kWh $R^2 = 0.30$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-11585.6271	3389.8445	-3.42	0.0007
Mean outside temp. (F)	158.1136	51.4000	3.08	0.0022
Median income (1000's)	-5.2501	9.3564	-0.56	0.5749
Household size	-560.8243	333.4175	-1.68	0.0931
% Owner occupied	47.1298	14.9000	3.16	0.0016
% Below povery	67.5361	18.9961	3.56	0.0004
Median home value ($1000's)	-1.6285	1.0040	-1.62	0.1054
Median age	35.4238	26.7674	1.32	0.1863
Median room count	1355.4969	292.1156	4.64	0.0000

Annual therms $R^2 = 0.38$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-29.3190	223.2142	-0.13	0.8956
Mean outside temp. (F)	-2.9126	3.1886	-0.91	0.3616
Median income (1000's)	0.6086	0.5505	1.11	0.2697
Household size	-27.0377	21.5213	-1.26	0.2098
% Owner occupied	-2.4219	1.1827	-2.05	0.0413
% Below povery	2.8270	1.4219	1.99	0.0475
Median home value ($1000's)	0.1109	0.0601	1.85	0.0657
Median age	3.6308	1.8968	1.91	0.0564
Median room count	124.8431	20.8445	5.99	0.0000

Table 2.1: Regression outcomes using household characteristics to explain annual electricity and natural gas consumption.

Naturally, some household characteristics, like income, home value, and owner occupancy are likely to co-vary with each other. To better quantify the independent contribution of each characteristic, we define and run linear regression models using the household characteristics to explain zip code mean annual electricity and natural gas usage. For the set S of characteristics, {outside temperature, median income, household size, % owner occupied, % below poverty, median home value, median age, median room count}, we run linear regressions defined by these equations for the ith home for all homes in a panel data set:

$$kWh_i = \alpha_i + \sum_{c \in S} \beta_{kwh,c} c_i + \varepsilon_i \tag{2.1}$$

$$therms_i = \alpha_i + \sum_{c \in S} \beta_{therms,c} c_i + \varepsilon_i \tag{2.2}$$

Table 2.1 on page 23 presents the fits for each of the models. The last column of values corresponds to the p-value probabilities that the null hypothesis, the coefficient in question is zero, is true. Smaller values represent higher confidence of a non-zero contribution from the corresponding coefficient. Coefficient names in bold correspond to p-values less than 0.05.

The model results generally re-affirm the qualitative results of the scatter plots, but elevated coefficient p-values also suggest significant covariation between the regressors and uncertain causal relationships. The R^2 values (0.30 and 0.38 for electricity and natural gas models respectively) indicate a lot of variation is left unexplained by the regression models.

As Figure 2.6 on page 22 suggests and the regression models confirm, there are statistically significant correlations between zip code aggregated population characteristics and energy use. It stands to reason that population characteristics can be used to quickly develop territory-wide targeting criteria or to steer efforts that are constrained by time or money. However, the variance explained by these regression models is modest, many modeled correlations are weak, and some relationships are counter-intuitive or highly sensitive to regression model formulation. For example, mean annual outside temperature was not found to have a statistically significant impact on natural gas usage; household size was found to be inversely related to consumption at 90% and 80% confidence thresholds for electricity and natural gas consumption, respectively; and home value is anti-correlated with electricity use at the 90% confidence threshold[4].

For all these reasons, demographic data used on its own can only be expected to achieve modest improvements in program targeting. Program planners and administrators should be careful not impose assumptions about the relationship between population characteristics and consumption that may not hold in practice. This is especially true now that more rigorous and accurate information can be derived through the empirical analysis of household meter data. We demonstrate in the remaining sections of this work that information about consumption patterns derived from meter data provide far more precise targeting criteria.

2.5.3 Thermal response

In the analysis of energy data, the most prominent driver of demand is often assumed to be thermal loads. However, 2.5.2 found no significant correlation between mean annual outside temperature and natural gas usage aggregated by zip code. In this section, we examine the impacts of climate and outside temperature on electricity and natural gas usage at the household level. For most California homes, residential cooling is electric, and heating is primarily fueled by natural gas, with modest electrical loads associated with heating system pumps and fans. We might therefore expect to see climate and temperature impacts on both energy types.

[4]One possible explanation for counter-intuitive relationships is the nature of real estate in California. Coastal cities tend to have more affluent residents, be more densely populated with smaller, urban residences, have a larger percentage of renters, and experience more moderate temperatures than inland cities.

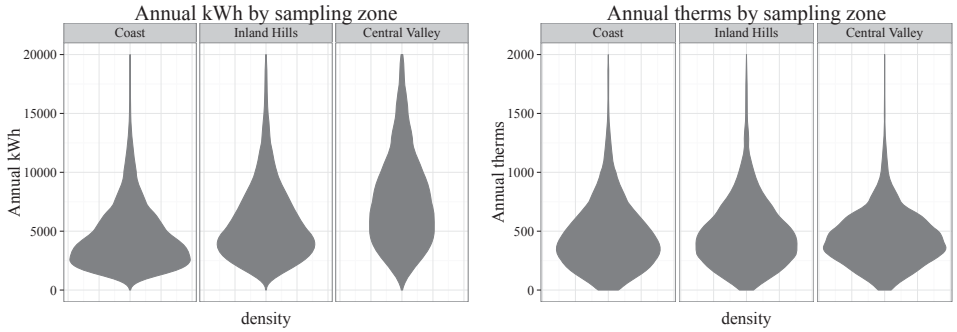


Figure 2.7: Distributions of annual energy by sampling zone, with electricity on the left and natural gas on the right. The Central Valley has the highest overall electricity usage, the coast the least, and the Inland Hills falls between the two. Natural gas consumption is more evenly distributes across sampling zones.

Cooling loads tend to follow the general pattern of increasing energy use with increasing outside temperature. Heating loads tend to increase with lower outside temperatures. Figure 2.7 on page 25 presents a very rough cut of the data that provides separate distributions of annual energy per geographic zone used to sample the data.

Distributions for electrical energy are on the left. It can be clearly seen that the cooler coast has the greatest number of lower-electricity-usage homes and the fewest higher-electricity-usage homes, while the Central Valley has the fewest lower-usage homes and the greatest number of higher-usage homes. This is consistent with a significant presence of cooling loads, but is most likely complimented by other structural or demographic factors. For example, land is less expensive in the Central Valley than the coast, so we can expect larger homes there.

Distributions for natural gas usage are on the right. The Central Valley has fewer high-usage residents than the other two zones. However, the differences among zones are not as significant as they were for electricity usage. This similarity is likely driven by several factors. First, a significant portion of natural gas usage goes to cooking, hot water, and clothes drying, which are ubiquitous end uses. Second, temperatures experience greater seasonal variation inland than on the coast. While inland locations may have fewer days requiring heating than the coast, some of those days will be much colder. Third, most of PG&E's service territory has very mild winters, so heating demand at levels common in other places with colder winters is only found in the tails of these distributions.

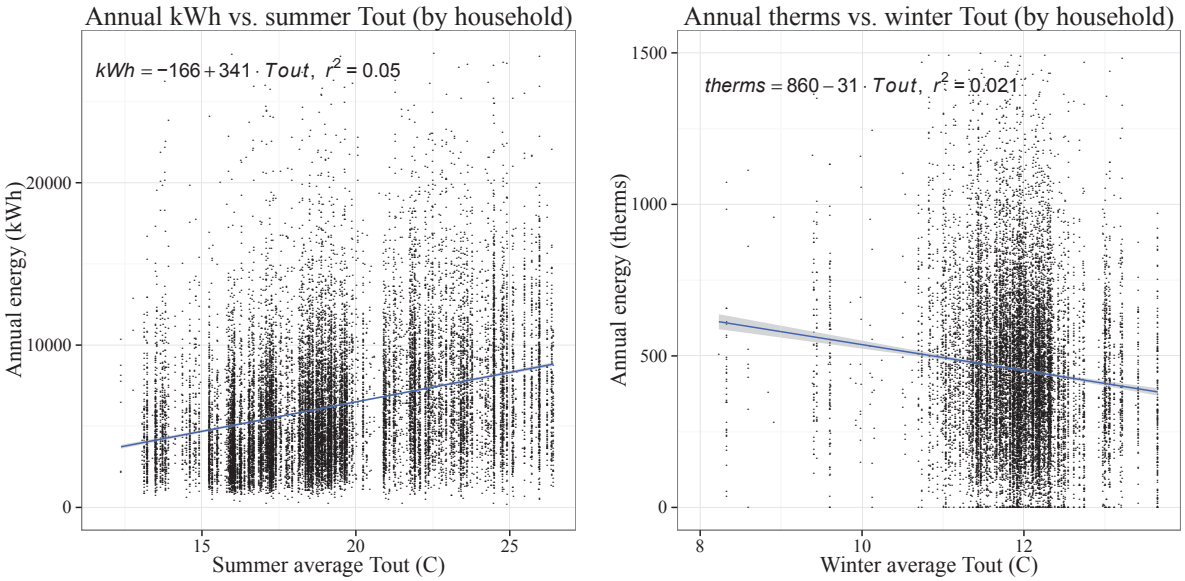


Figure 2.8: Household values for annual electricity usage vs. mean summer temperature (right) and annual natural gas usage vs. mean winter temperature (right). Here winter is defined as November through April and summer is defined as May through October. There is one data point per household.

The magnitude of climate impact on heating and cooling loads is most simply illustrated using scatter plots of annual energy vs. outside temperature. Figure 2.8 on page 26 provides one scatter plot for annual electricity usage and another for annual natural gas usage. Both plots have one data point per household, with annual energy on the y-axis and mean seasonal temperature on the x-axis. The plot on the left shows annual electricity usage against mean summer temperatures. The plot on the right shows annual natural gas usage against mean winter temperatures. A simple regression fit is also provided for each to more clearly illustrate and quantify the best fit for the temperature dependence, in terms of annual kWh or therms per degree of average temperature increase.

The temperature dependence is statistically significant for both types of energy, but within specified temperature ranges, there is very substantial scatter in energy use. The model R^2 values indicate that seasonal temperatures explain just a small portion of observed variation in annual energy consumption. All the other factors that contribute to energy consumption — home size, the preferences of occupants, equipment ownership, and operational details — are presumably responsible for the rest of the variation.

This observation partially undermines (at least in CA) the premise of many energy models that residential energy is best explained by outside temperature interacting with heating and cooling equipment. However, some homes are clearly more influenced by outside temperatures than others. Approaches capable of disaggregating the thermal and non-thermal components of energy use can provide distributions of heating and cooling

consumption that better differentiate between low and high consumption. In many cases, the higher-consumption households will make better targets for attention from utility programs.

2.5.3.1 Disaggregating natural gas usage

Natural gas is used for cooking, hot water, drying clothes, and space heat. Except for space heating, all of these uses can be expected to be relatively consistent throughout the year[5]. Space heating should increase in colder weather and will typically cut off entirely in warmer weather. For these reasons, warm weather natural gas usage should be restricted to the consistent demands of hot water, cooking, and clothes drying, and the difference between warm weather consumption and cold weather consumption should approximate the energy use of space heating.

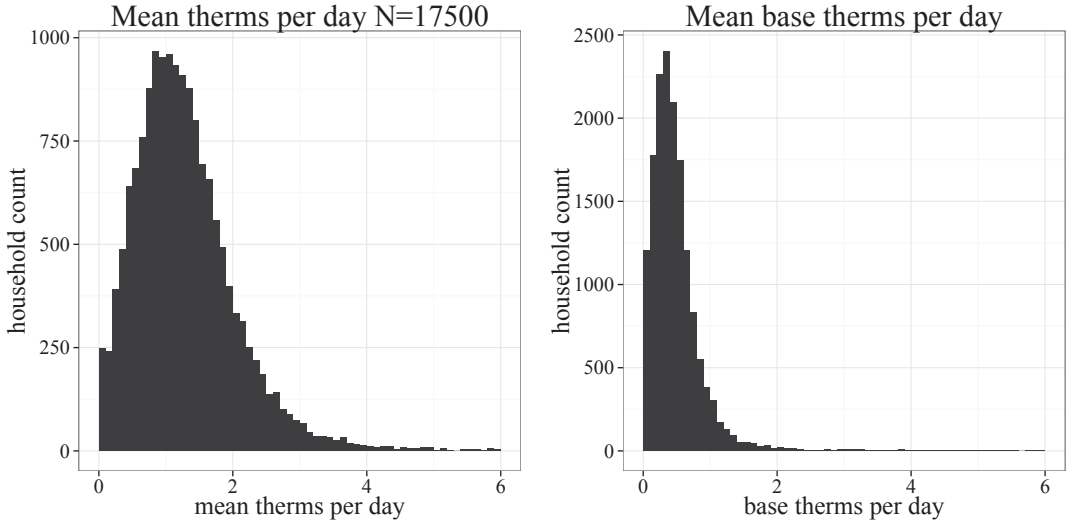


Figure 2.9: Daily mean total and base therms, where base therms include water heating, cooking, and any other loads that do not vary seasonally

In equation 2.3, we define a regression model designed to identify heating and base therm usage using daily observations of therm usage and mean daily outside temperature, T_{out} in degrees Fahrenheit. To account for the threshold between warmer weather without heating and cooler weather with heating, we define T^{-}_{out} as $max(0, 65 - T_{out})$, which measures the degrees the daily mean temperature is below 65°F.

$$therms_d = \alpha + \beta T^{-}_{out,d} + \varepsilon_d \tag{2.3}$$

[5]Water heating will also vary with seasonal changes in inlet water temperature and the ambient temperature of air around the water heater, but these effects should be modest compared to space heating demands.

This model estimates base consumption as α therms per day at 65°F. Annual heating therms can be estimated as $\sum_{d \in days} \beta T^{-}_{out,d}$, or as the difference between the observed total and the daily based usage times the number of days. For convenience, we use the latter method. Figure 2.9 on page 27 presents histograms of annual total (left) and estimated base (right) therm consumption for the portion of our data sample with at least one year worth of gas usage data.

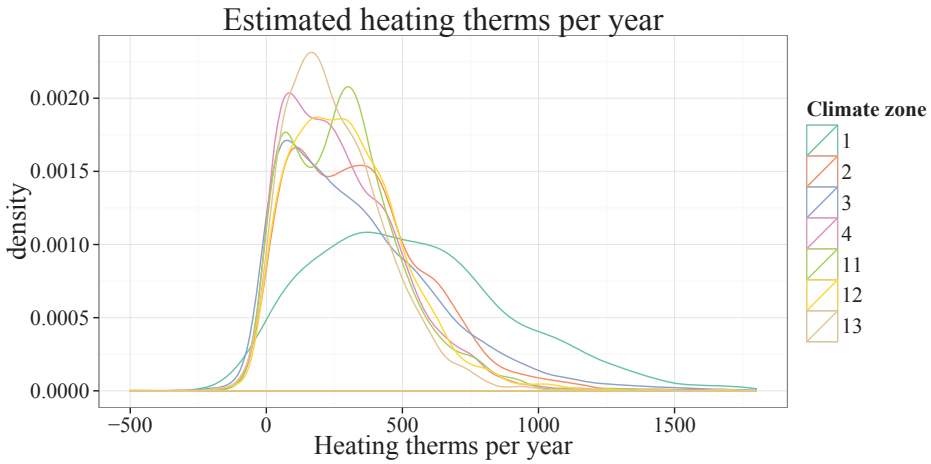


Figure 2.10: Estimated therms used annually for heating across all households in the sample broken out by California climate zone.

Figure 2.10 on page 28 presents density curves — one per climate zone — for annual heating demand of the subset of our sample with at least a year of natural gas usage data, as estimated by our regression model. Because the demand is a regression estimate of therm usage as a function of outside temperature, and some homes have consumption uncorrelated with temperatures, a handful of values are below zero. Zones 1-4 progress from north to south along the cool coast and zones 11-13 progress from north to south in the hotter Central Valley. Zone 1 is the coolest and also the zone with the highest estimated usage. The other climate zones have lower estimates overall. The climate responsive nature of this result suggests that the regression model did a good job separating out the heating loads. The significant overlap of all the distributions indicates that differences in consumption within climate zones are greater than across climate zones. This means that targeting criteria that make use of individual household heating consumption estimates will significantly out-perform coarse targeting by climate zone. Several of the distributions appear to have a bimodal shape to them. These may be due to differences in appliance ownership, possibly electric vs. natural gas hot water or central gas furnaces vs. room heaters or electric heat.

2.5.3.2 Disaggregating electricity usage

The same basic technique for disaggregating temperature responsive and non-responsive loads can be repeated using electricity consumption data. In equation 2.4, which is analogous to equation (2.3), we define a regression model designed to identify air conditioning and base electricity usage using daily observations of electricity usage and mean daily outside temperature, T_{out} in degrees Fahrenheit. To account for the threshold between colder weather without cooling and hotter weather with cooling, we define T^{+}_{out} as $max(0, T_{out} - 65)$, which measures the degrees the daily mean temperature is above 65°F.

$$kWh_d = \alpha + \beta T^{+}_{out,d} + \varepsilon_d \tag{2.4}$$

This model estimates base consumption as α kWh per day at 65°F. Annual electric cooling energy can be estimated as $\sum_{d \in days} \beta T^{+}_{out,d}$, or as the difference between the observed total and the daily based usage times the number of days. For convenience, we use the latter method.

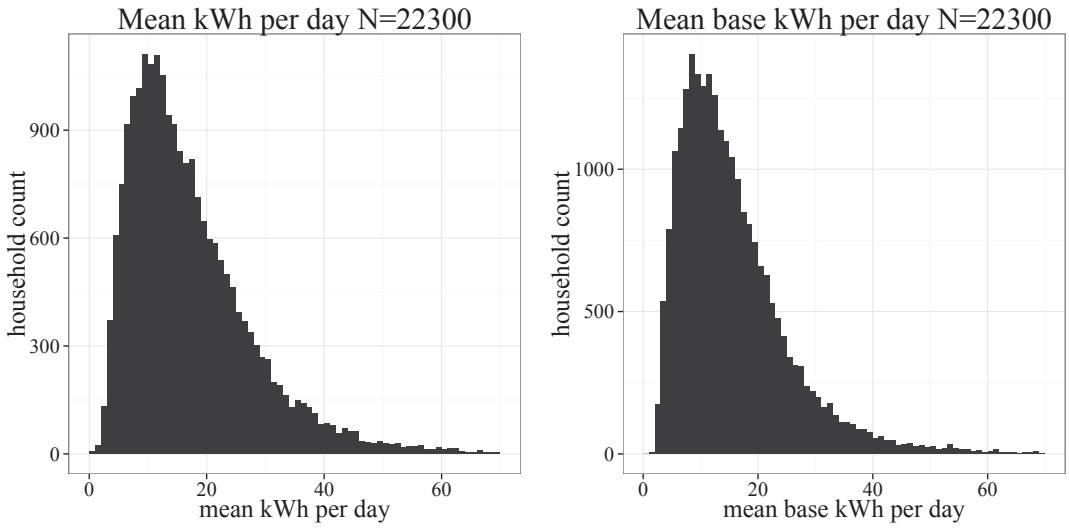


Figure 2.11: Daily mean total and non-thermal base kWh. Cooling consumption is a modest fraction of total energy, so the distributions look quite similar.

Figure 2.11 on page 29 presents histograms of annual total (left) and estimated base (right) electricity consumption for the portion of our data sample with at least one year of electricity usage data. The distributions are quite similar, which is consistent with the observation that many California homes, especially homes along the coast, do not require air conditioning. Even in homes with air conditioning, the cooling energy tends to be a small fraction of total electric energy use in homes.

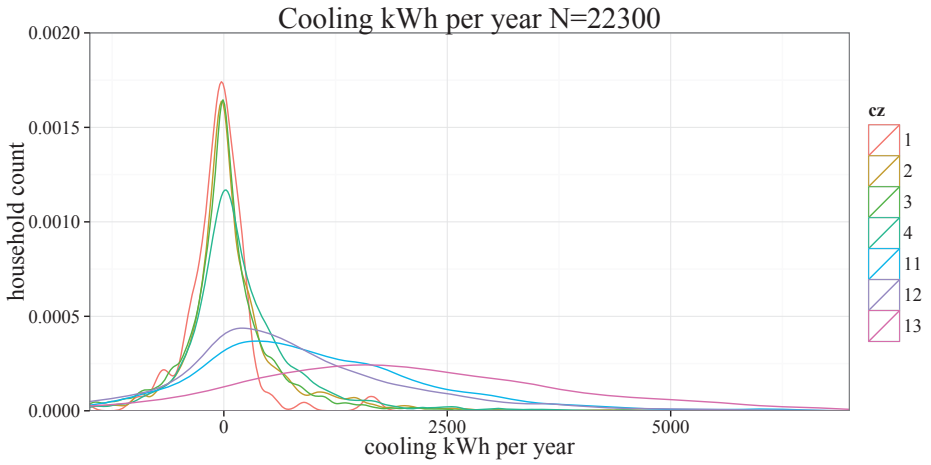


Figure 2.12: Estimated kWh used annually for cooling across all households in the sample broken out by California climate zone.

Figure 2.12 on page 30 presents density plots for annual cooling energy by California climate zone. Zones 1-4 progress from north to south along the cool coast and zones 11-13 progress from north to south in the hotter Central Valley. Zone 13 is the hottest and also the zone with the highest estimated usage. The next highest usage comes from zones 12 and 11, which are the second and third hottest zones and are also in the Central Valley. The remaining coastal climate zones have distributions that are close to mean zero, suggesting that most home go without air conditioning and the observed variation is largely due to model errors. This plot makes clear where the high cooling demand is within California, but again the breadth of the distributions suggests that targeting specific high consuming homes within each zone will prove superior to targeting by climate zone.

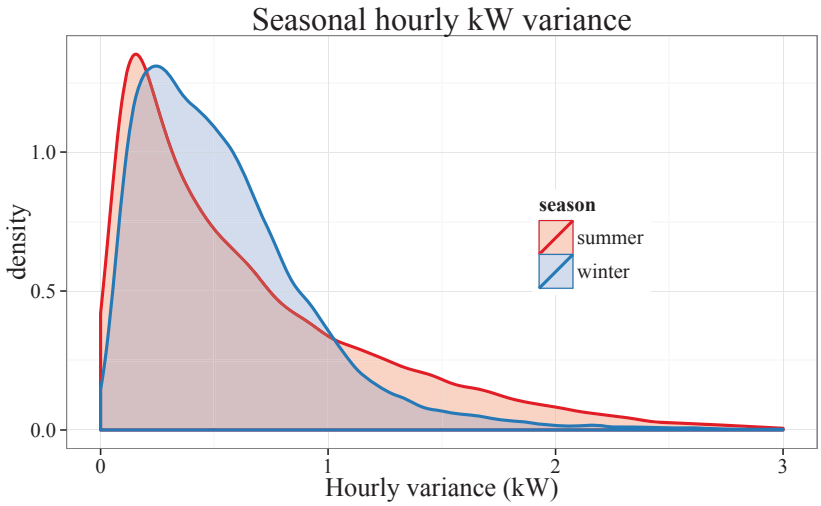


Figure 2.13: Difference in hourly variance between summer (May-October) and winter (November-April) loads. In general, hourly variance is higher in the summer and lower in the winter. This is consistent with a seasonal use of air conditioning that increases variance when operating. The magnitude of the variance is also consistent with the power draw of central air conditioning units.

Another signature of cooling loads should be their seasonal nature. During periods when cooling is inactive, all the background variance in energy use observed can be attributed to those other factors. However, when cooling is active, it tends to cycle on and off according to control inputs. This creates additional variance distinct from the background. Figure 2.13 on page 31 provides a look at the distribution of variance in hourly energy consumption for just the summer months (May-October) and just the winter months (November-April) across all the buildings in the sample. The differences are obvious and striking. First, as expected, the summer period has a larger proportion of homes with high variance than the winter period. At the same time, the summer variance in some buildings is lower than winter. These homes almost certainly do not have air conditioning or are shut down in the summer due to seasonal occupancy. It is worth noting that seasonal patterns of lighting demand or occupancy for homes can also produce changes in variance. For example, a ski house (there are many in the Sierra Nevada mountains) may sit mostly unused for the summer but be used far more actively in the winter, especially during weekends or vacation periods. In principle, such patterns ought to be distinguishable from those caused by thermal loads.

2.5.4 Timing of use

Everyday experience suggests that residences, just like their inhabitants, should exhibit daily diurnal patterns of demand, typically lower overnight and highest during periods

of consistent and active occupancy. Evenings will tend to be the times when occupants return home from daily activities and use electricity for lighting, space conditioning, and other activities. Because the timing of demand is an important design criteria for grid planning and reliability, we turn our attention in this section to the question of when households reach their peak electricity consumption.

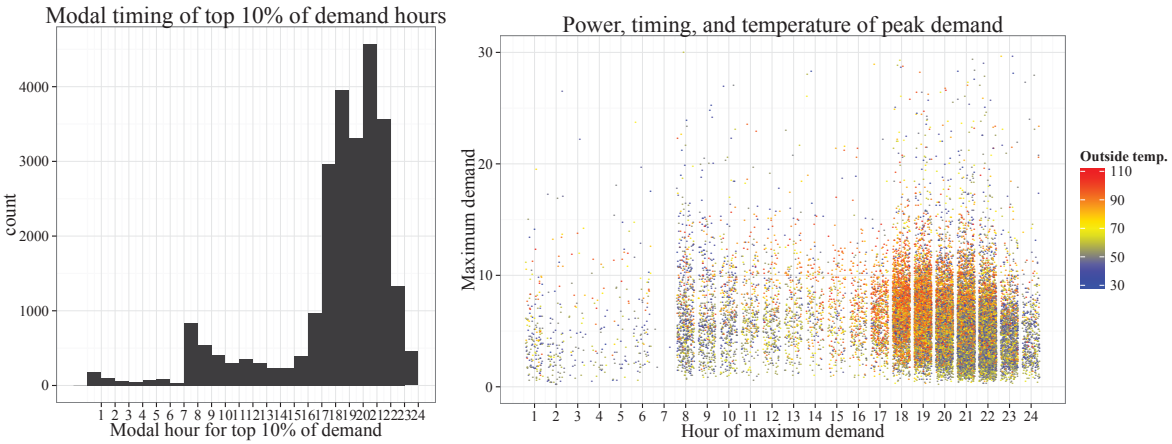


Figure 2.14: Histogram of the modal hour of day for the top 10% of demand hours for each household (left) and plot of hour and magnitude of peak demand for each household, colored by the outside temperature at the time (right).

Figure 2.14 on page 32 examines periods of peak energy use as indicators of maximal activity in homes, presumed to be a combination of automated loads, like heating and cooling, and occupant driven loads, like lights and appliances. On the left side is a histogram of the modal timing of the top 10% of hours from each home. On the right side is a plot with the hour of day and magnitude of each household's top hour of demand, colored by the outside temperature at the time. Both metrics have clear evening peaks, with increasing frequency in the early afternoon and very low values during the early hours of the morning. The color coded peak hour plot suggests that warmer temperature peaks are concentrated in the afternoons, while cooler temperature peaks can be found in the mornings, afternoons, and evenings. This pattern is consistent with the expected patterns of occupancy-driven loads and scheduled heating and cooling loads.

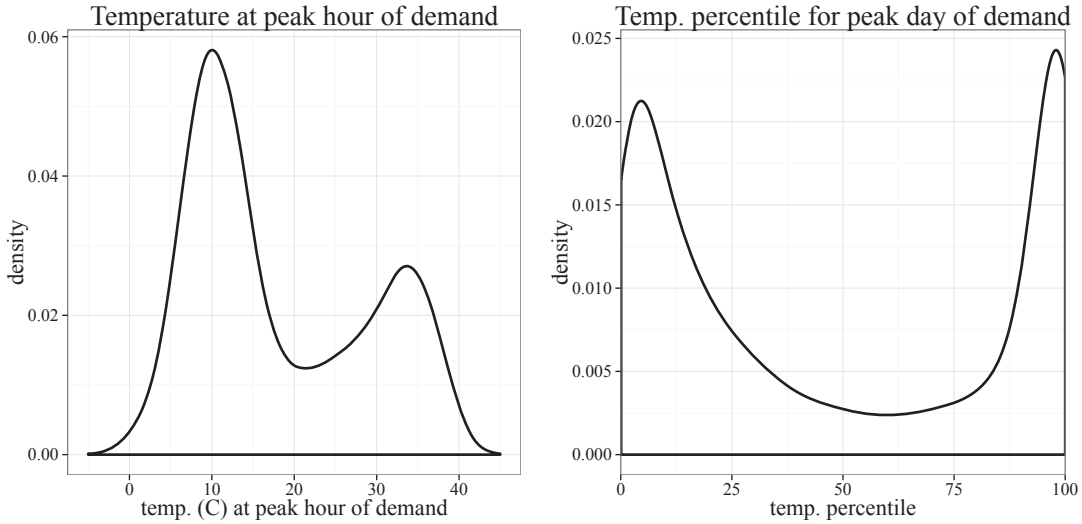


Figure 2.15: Temperature at the peak hour of demand and percentile of temperature for the peak day of demand

To further examine the relationship between temperature and peak demand, Figure 2.15 on page 33 provides the distribution of temperatures during the peak hour for each home in the sample (left) and the percentile of the average temperature during the day of highest total demand (right). Both distributions are strongly bimodal and suggest a simple classification of buildings into high temperature peaking and low temperature peaking categories. It is probable that homes without air conditioning will peak on a cold winter day due to auxiliary heaters, pumps, and fans as well as longer periods of lighting and increased probability of residents at home and indoors. Similarly, it is probable that homes without electric heat and with air conditioning will peak in the summer.

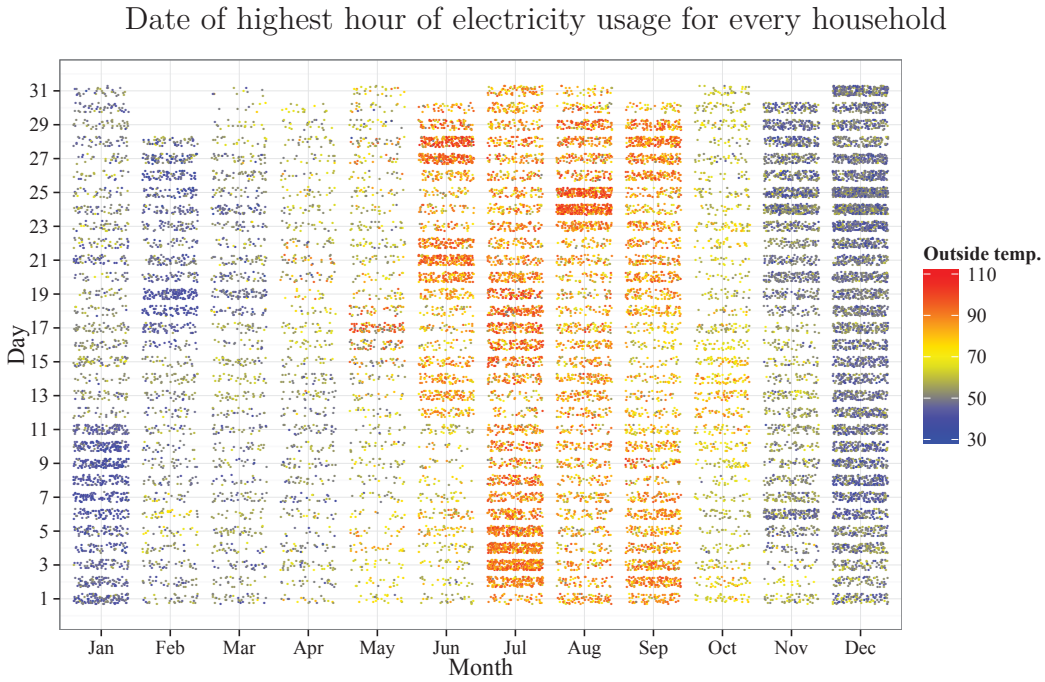


Figure 2.16: Jitter plot of month of year and day of month of peak hour of demand for every household, colored by the coincident outside temperature.

Figure 2.16 on page 34 is a plot of month of year and day of month for the peak hour of demand for every household. It is colored by the coincident outside temperature and both x and y values have added random noise to spread out the overlapping points. This calendar view of peak demand clearly illustrates the seasonal nature of peak demand, with the highest frequency days occurring during the hottest part of the summer (Jun-Sep) or the coldest part of the winter (Nov-Feb). This confirms that thermal loads play a significant role in peak demand days. In the summer months, sequential days of high consumption, very likely caused by heat waves, are evident. In the winter, the peak days are clustered around the end of the year, suggesting that holiday gatherings, cold weather, and long, dark, nights combine to produce peak demand in many homes. This view of the data allows for efficient identification of the homes that make the best targets for cooling efficiency or demand response programs seeking to reduce summer peak loads.

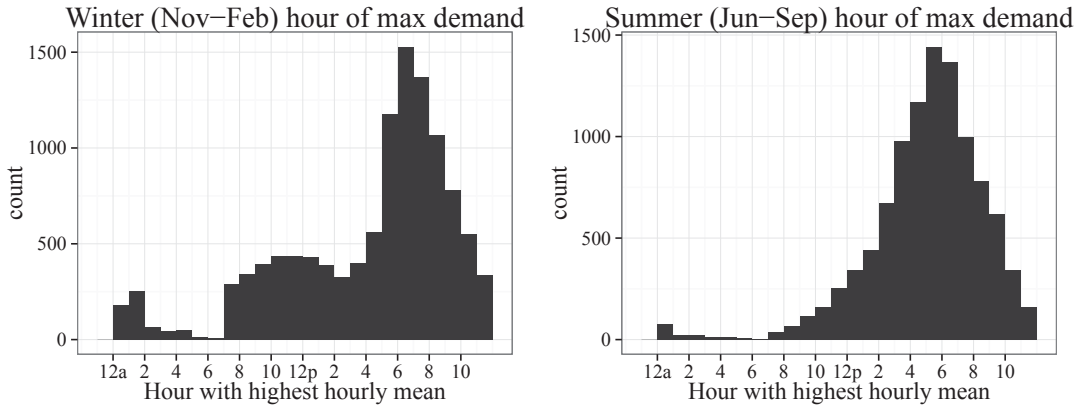


Figure 2.17: Comparison on the timing of peak usage for winter peaking and summer peaking homes.

Based on the months with the highest incidence of peak demand hours for homes, the data can be separated into winter peaking and summer peaking subsets of homes. Figure 2.17 on page 35 shows histograms of the hour of day of peak demand for each of these subsets. Clearly the character of usage is different between the two groups. The winter histogram suggests the majority of peaks happen in the early evening, but with significant numbers of homes with mid-day peaks as well. The summer histogram peaks in the early evening as well, but with smooth ramps on either side. The smoothness of the summer histogram is most likely due to the likelihood of peaking due to air conditioning loads in the afternoon superimposed on patterns of occupancy that tend to increase occupant-driven demand in the early evening.

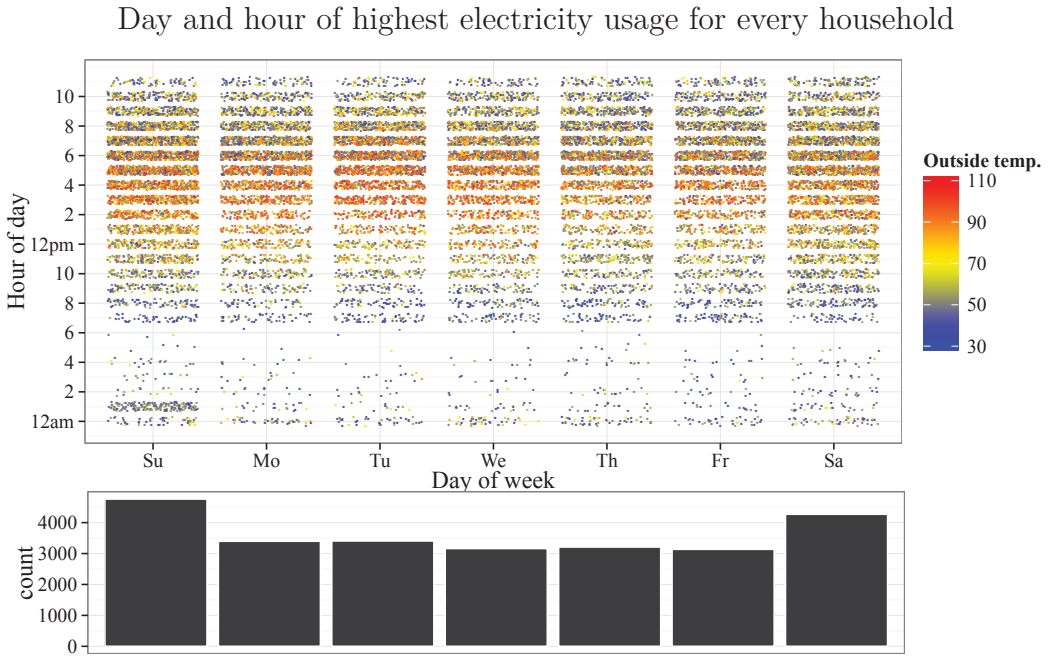


Figure 2.18: Jitter plot of the day of week and hour of day of the peak hour of demand for each household, colored by the peak coincident outside temperature, with a histogram of counts of peak hours per day of week at the bottom.

Figure 2.18 on page 36 is a jitter plot of the day of week and hour of day of the peak hour of demand for each household, colored by the peak coincident outside temperature. This view of the data reveals differences in the counts of peak hours of demand by day of week. Each weekend day is more likely to witness peak demand than each weekday. This is likely due to increased occupancy on weekends, presumably due to most people being off from work and school on weekends. More subtly, it seems that peak demand is more likely toward the beginning of the week than the end. This might be due to patterns of chores involving large appliances, and it might also be due to occupancy, with households more likely to eat out or otherwise spend time outside the home later in the week. Regardless of the specific reasons, the weekly patterns of peak demand strongly support the idea that occupancy and occupant preferences must be accounted for alongside other factors, like thermal conditions and demographics, to understand patterns of energy use.

summer peak (Jun-Sep)	40%
evening peak (5-10)	55%
peak > 5kW	59%
summer evening peak	22%
summer evening peak > 5kW	16%
winter peak (Nov-Feb)	45%
daytime peak (8am-3pm)	30%
daytime winter peak	12%

Table 2.2: Percentage of homes that fall into various load peaking categories.

Using patterns in peak timing presented in this section, groups of residences with desirable properties for demand response or efficiency program participation can be readily identified. This data allows homes to be targeted by using the season, day-of-week, day-of-month, and hour-of-day of peak demand. The magnitude of demand and outside temperature during the peak can also be factored into targeting rules. Table 2.2 on page 37 provides some breakdowns of the percentage of homes in various peak demand categories.

2.5.5 Load shape metrics

Load shape metrics are values computed using simple heuristics that capture basic features of daily load shapes [91]. Related metrics can also capture thermal response [11]. These metrics, described in Table 2.3 on page 38, capture base load, peak demand, mean demand, and variability, each of which is valuable information to grid operators. Furthermore, combinations of these metrics suggest prevailing efficiency- and demand response-relevant characteristics of home operation. For example, the efficiency opportunities in a high-mean, low-variability home (i.e., reducing baseload and introducing better controls to shut off unused equipment) are different than those in a high-mean, high-variability home (i.e., increasing the duration of low-demand periods and reducing the magnitudes of peaks).

Load shape parameters

Load shape parameter	**Description**
min	The 1st percentile of power demand for the day
max	The 99th percentile of power demand for the day
range	max - min
mean	The mean of all power demand for the day
min/max	The ratio of min and max
night/day	The ratio of usage from 2 to 5am and 2 to 5pm

Annotated load shape

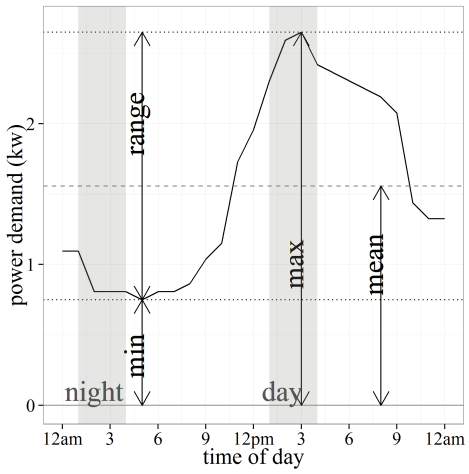


Table 2.3: Description of load shape metrics with labeled schematic of a building load curve. The x-axis spans a 24-hour period; the y-axis is whole building power demand. The minimum, or base, demand is unlikely to be zero.

Trends in specific metrics can help reveal changes in load patterns over time. For example, an increasing daily minimum value might signal the accumulation of vampire loads[6] from new electronics or the addition of a second freezer to a home. A relatively high night-to-day ratio might indicate that the occupants forgot to turn off the porch light before going to bed. When averaged across all available days, load shape metrics distill important operational details of buildings into a compact set of numerical values suitable for comparison across populations of buildings. They are similar in spirit to distributions of the energy used by various end uses [87], but they do not require any additional metering. Figure 2.19 on page 39 provides distributions for six different load shape metrics drawn from the entire sample of residences studied. Building upon benchmarking methods based on annual energy consumption, for example [75, 32], the values of each residence's load shape metrics can be placed into multiple distributions of full population (or peer group subset) metrics that can triangulate specific operational practices relevant to efficiency or demand response program goals.

[6]Vampire loads are the power drawn by the standby and off states of electronics. They can be minimized by good design.

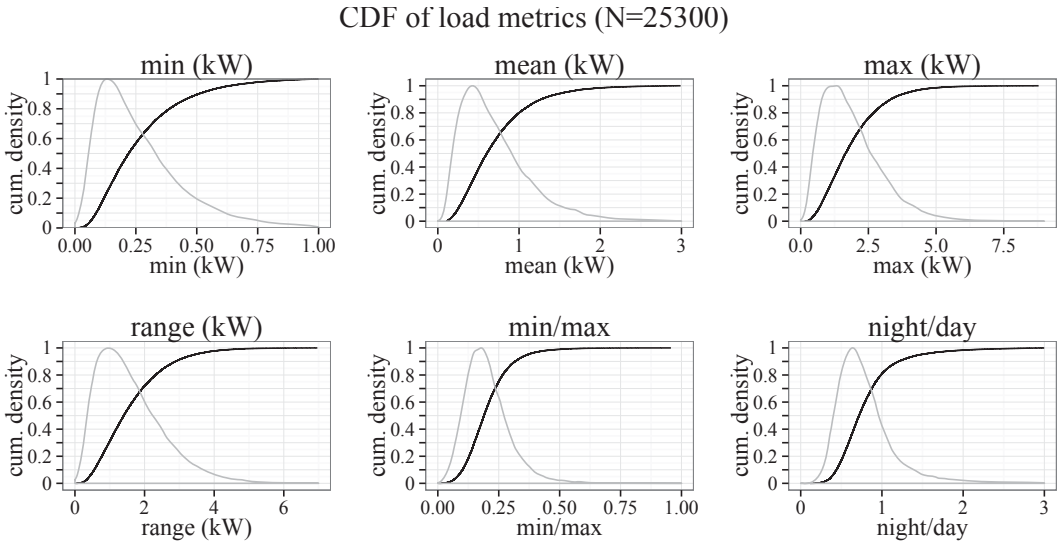


Figure 2.19: Empirical cumulative distributions of load metrics across all residences, with the related density curves in gray (and re-scaled) for reference.

2.5.5.1 Base load

Historically, investments in improving the efficiency of refrigerators has paid off multiples on the investment [85]. An efficiency program focused on turning off or replacing old power-hungry refrigerators might target the subset of the population with relatively high minimum usage. With the proliferation of digital and other electronics, plug loads have been identified as the fastest-growing domestic end use. The power demands of poorly design electronics are often substantially larger than well-designed products that perform the same function. The designers of appliance standards and efficiency programs might rely on the 'min' and 'min/max' metrics to better understand the potential savings associated with equipment replacement. Moreover, the values of these metrics could be used to identify specific households likely to benefit from a reduction in their always-on loads.

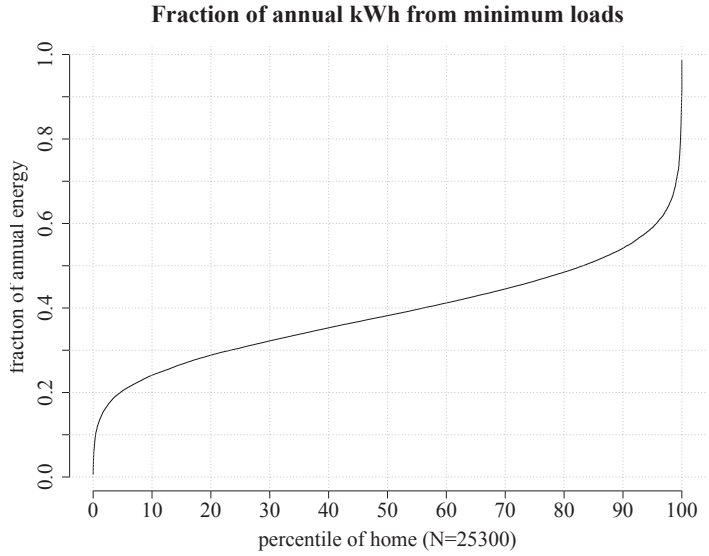


Figure 2.20: Contribution of apparent baseload, based on the 'min' metric, to annual energy. This is an indication of how big a contribution 24x7 loads make to energy consumption.

Using the overnight minimum demand as the *apparent baseload*, we multiply by 24 hours to calculate the total energy consumed per day powering just the baseload. By running this calculation for every day for every residence, we calculate the fraction of each residence's annual energy dedicated to baseload. These values, sorted from smallest to largest, are presented in Figure 2.20 on page 40. This plot demonstrates that apparent baseload is less than 20% of annual energy use in fewer than 5% of homes and is at least 40% of annual energy usage for 45% of all homes.

These results strongly suggest the importance of improved controls, more efficient always-on appliances, and better standby and off states for electronics. These goals have traditionally been pursued through a combination of standards, mandates, and efficiency program incentives. Each of these can be better informed by the availability of related metrics at the household level. Baseload changes resulting from programs and standards are also among the easiest impacts to measure. The min and mean load shape metrics should capture any sustained improvements.

2.5.6 Demand during system peak

Grid planning revolves around the simple fact that, in electric power systems, supply must meet demand all the time and everywhere. A significant portion of expenditures on infrastructure, which tend to run into the billions of dollars, are dedicated to ensuring that systems have the capacity to serve their peak expected demand. These costs buy redundancies in generation, transmission, and distribution infrastructure and are typically

non-negotiable because when the planned usage is exceeded, the grid will fail, causing costly and destructive outages and public outrage.

Today's planners have new worries. Climate change mitigation is likely to require the integration of large quantities of intermittent renewable energy generation onto the grid. It will be very expensive to solve these new problems with redundant infrastructure. Attention is turning to the feasibility of more actively shaping demand to reduce grid congestion and avoid infrastructure investments.

Demand response programs are designed to encourage load curtailment by utility customers during periods of grid stress. This section develops a simple use case for targeting demand response program enrollment using smart meter data.

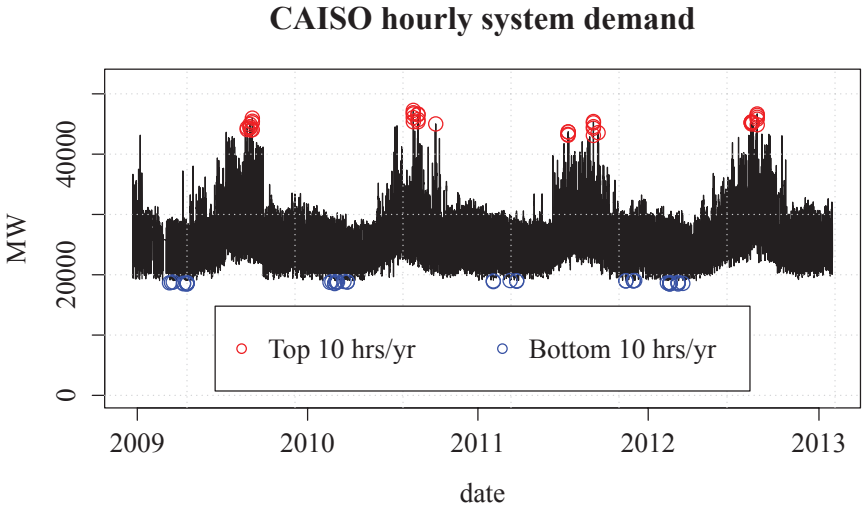


Figure 2.21: CA ISO demand from 2009 to 2012, with annual top and bottom 10 hours highlighted.

Figure 2.21 on page 41 is a time series plot of California's hourly total system demand between 2009 and 2012, inclusive. The highest 10 hours of demand for each year are highlighted with red circles and the lowest 10 hours of demand are highlighted with blue circles. For reference, eliminating the top 10 annual hours of demand would reduce the multi-year system peak by 8.8%, and eliminating the top 20 hours would reduce the peak by 10%. The peak observations tend to come in the afternoon or early evening of the hottest days of the year. The lowest observations tend to come overnight in mid-winter or early spring.

We locate the peak coincident demand for each household by looking up all the meter readings that correspond to the hours of system peak demand. We then calculate *household average peak coincident demand* by averaging peak coincident demand for each household. We repeat this process with the minimum hours of system demand. These

household average peak and minimum coincident demand values (one of each per household) can then be used to study the distribution of demand during system peaks.

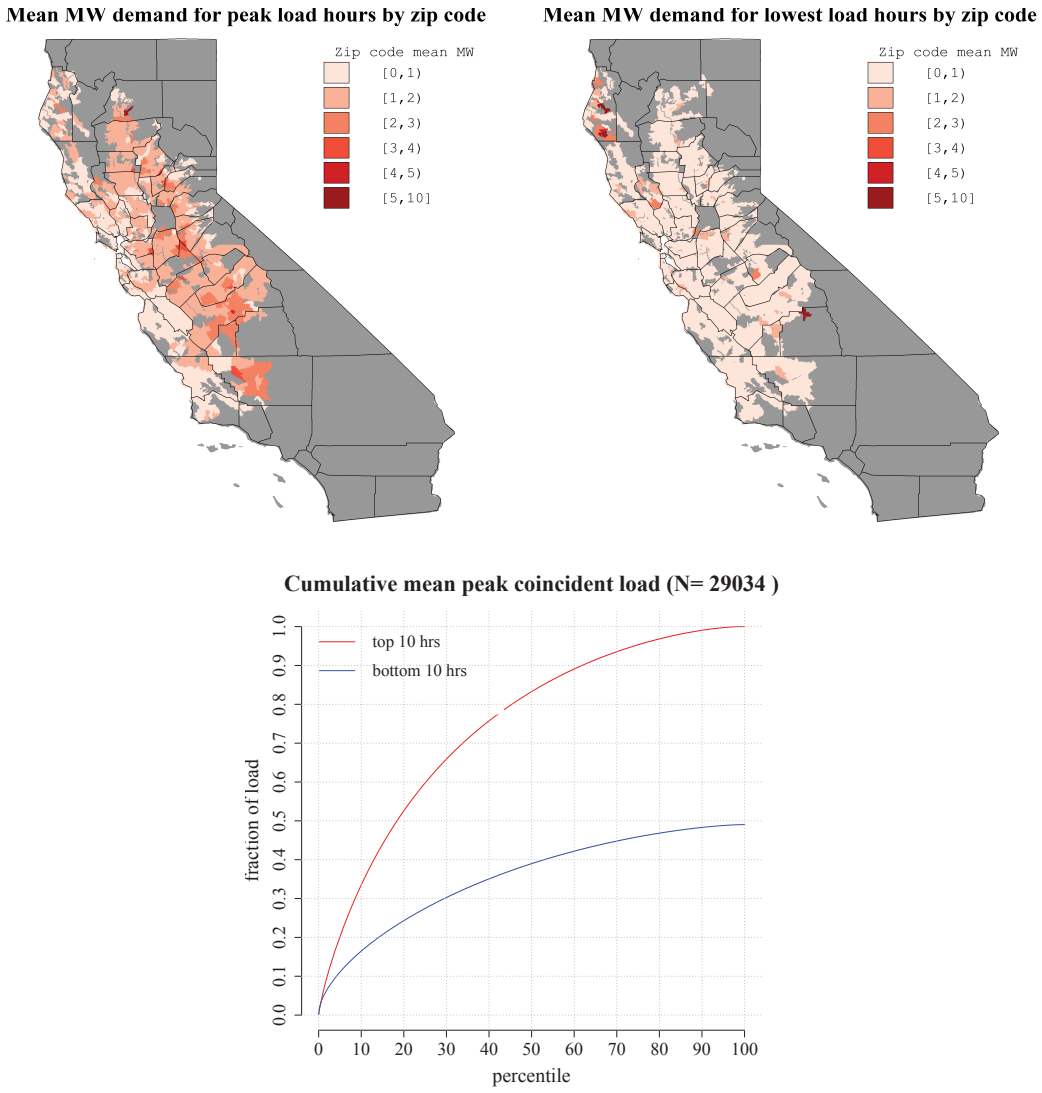


Figure 2.22: Cumulative sum across households of peak and minimum coincident demand, sorted from highest household to lowest (bottom center) and maps of household average peak and minimum coincident demand for annual top 10 peak hours (upper left) and annual bottom 10 hours (upper right) from 2009 to 2012 averaged by zip code.

Figure 2.22 on page 42 presents the results of our peak and minimum coincident demand calculations. On the bottom is a cumulative distribution of each. This plot shows that on average, just 10% of households are responsible for nearly 35% of total residential load during peak demand periods and 20% of household account for just over 50% of total residential load. The lowest-consuming 50% of households are responsible for

just over 15% of total residential demand. The unevenness of the distribution of demand during peak system demand periods is underscored by the flat distribution of demand during minimum system demand periods. A random sample from the top 10% of peak coincident demand is expected to contribute 3.3x the demand of a sample of similar size drawn from the general population.

The map of zip code mean peak coincident demand on the upper left of the figure provides a spatial view of where demand is coming from during peak periods. It is readily apparent from this view that the Central Valley has systematically higher demand than the Coastal or Hills regions. This affirms the expectation that cooling loads drive the peaks because the Central Valley routinely reaches temperatures approaching 40°C (104°F) on summer afternoons.

The map of zip code mean minimum coincident demand on the upper right of the figure uses the same color scale as the maximum demand map to display the geography of demand during system minimums. Most of the territory is low usage, but there are pockets of higher use, possibly associated with electric heating or unusually high lighting loads.

Grid transmission congestion can also be a significant problem on high usage days. The spatial nature of the information contained in this result suggests another grid management application. Residential demand data, located using zip codes or utility specific knowledge of which distribution infrastructure serves which customers, could be combined with location-sensitive congestion information to determine where the largest residential contributors to congestion are located.

Taken all together, the skewed distribution and clear spatial pattern of demand during peak periods strongly reinforces the idea of targeting demand response and efficiency measures, most likely focused on cooling demand, to achieve more reliable and cost-effective demand reductions during peak periods and to reduce expenses associated with providing a grid operating margin for peak demand periods.

2.6 Discussion

This work has focused on methods of differentiating patterns of household energy use to better target efficiency and demand response programs. We have focused on methods and metrics that are accessible and relatively easy to compute in the hope that they will be adapted and used by real-world programs. However, there is certainly room for improvement in much of what we have done. In some cases, the analysis could be even simpler without losing accuracy. In others, different methods might improve upon the performance of our methods.

We acknowledge that the regression models we used to disaggregate natural gas and electricity loads into thermal and non-thermal components are potentially biased by the use of a fixed threshold temperature of 65°F and that other regressors would likely help to better explain the variance in the data. In fact, we have completed a detailed treatment

of regression models for daily kWh that tackles these modeling issues in Chapter 3.

Higher potential for savings associated with the top residences according to various metrics could be deflated by systematic differences in those residences. For example, peak coincident load is highest in the hottest regions of the state. It could be argued that those residences have a unique need for more cooling than the rest of the population. Singling them out for demand response could be asking the least comfortable households to be even less comfortable.

The single peak hour of household demand used characterizing the timing of peak residential loads in 2.5.4 could also prove sensitive to factors of little interest to utility program implementers. A more robust metric would likely span several of the top hours to control for fluke occurrences. We have relied upon our sample size to discern patterns in those single hour peaks, but targeting should most likely be done with a more robust metric.

Finally, the most obvious shortcoming of this work is that it is speculative in nature. We have not been granted access to program design or implementation processes, and therefore have not been able to evaluate the efficacy of our methods under controlled conditions. We would expect some surprises when these methods do get tested in the real world.

However, even with the limitations we have listed, the methods presented here represent a dramatic improvement over standard targeting criteria, like annual usage or demographics on their own. It would be useful to develop a data set with some ground-truth professional efficiency evaluations associated with homes to allow for more precise quantification of targeting accuracy.

2.7 Conclusions

Depending on program objectives, there are many different ways to use smart meter data to target household for participation in energy efficiency or demand response programs. Our analysis focused on several categories of targeting metrics and consistently found skewed distributions and other criteria suitable for improving the targeting of programs through the use of household level consumption data. The information derived from interval meter data is sufficient to dramatically improve outcomes over programs with self-selected or even loosely targeted participation.

2.7.1 Annual usage

Distributions of annual electricity and natural gas usage by household are highly skewed, with fat tails extending toward higher usage. Skewed distributions are indicative of potential gains achievable through effective targeting. The top 10% of homes use approximately 25% of total electricity. The lowest 50% of homes use just over 25% of the total electricity. Similar skew is present in the natural gas data. The mean consumption

of the top 10% is 2.5x the sample average, so targeting a program that achieves savings proportional to total consumption would be expected to yield 2.5x the savings of an untargeted program. Annual electricity and natural gas usage are correlated, but only loosely. Homes with similar or dramatically different usage of the two energy sources can be readily identified.

2.7.2 Household characteristics

Energy usage follows general patterns that are partially explained by local home and demographic characteristics. In our sample, the percentage of owner occupied homes and the mean number of rooms correlated with both electricity and natural gas consumption when aggregated by zip code. However, median income, home value, the percentage of people below poverty, and the median age of residents correlated more with natural gas consumption than electricity consumption. The size of households did not correlate with either. The R^2 values for regression models of zip code mean consumption of electricity and natural gas against outside temperature and 7 demographic variables were only 0.30 and 0.38 respectively. Wherever possible, targeting efforts should be based on observed patterns of consumption in addition to demographic data.

2.7.3 Thermal response

Correlations between average outside temperatures and household level annual electricity and natural gas consumption are present in the data. However, the variability across households within narrow temperature ranges is far greater than the variability across temperatures. Smart meter data allows regression models to resolve the temperature sensitivities at the individual household level. Comparisons of temperature sensitive loads across climate zones demonstrate that the variability within each zone is far greater that the variability across zones. By providing information about individual households, these methods can support more accurate targeting within climate zones, zip codes, or demographic categories.

In the case of natural gas, usage can be broken down into base consumption — which includes hot water, cooking, and clothes drying — and space heating. Interventions can be developed separately for homes with high base usage and homes with significant heating loads. Electricity use can be broken down and targeted in a similar manner, with different interventions designed to address base and cooling loads.

2.7.4 Timing of peak

Peak electricity demand timing was revealed to fall in a bimodal distribution between winter months and summer months, roughly correlated with highest and lowest temperatures of the year. Weekend days see more residential demand peaks than weekdays, and both winter- and summer-peaking households tend to peak in the early evening. Using

patterns in peak timing presented in this section, groups of residences with desirable properties for demand response or efficiency program participation could be readily identified. This data allows homes to be targeted by using the season, day of week or month, and hour of day of peak demand. The magnitude of demand and outside temperature during the peak can also be factored into targeting rules. For example, 16% of our sample homes experienced peaks $>$ 5kW between June and September and between 5 and 10pm. These homes are very good candidates for enrollment in a demand response program.

2.7.5 Load shape metrics

Extending benchmarking practices that make use of monthly or annual energy totals, the values of each residence's load shape metrics can be placed into multiple distributions of full population (or peer-group subset) metrics that can triangulate specific operational practices relevant to efficiency or demand response program goals. For example, *apparent baseload* is a metric of total annual energy consumed by loads maintained at each home's daily overnight minimums. Apparent baseload is responsible for at least 40% of annual energy usage for 45% of all homes, and it is less than 20% of annual energy use in fewer than 5% of homes. Plug loads and always-on appliances should clearly remain targets of future efficiency programs.

2.7.6 System peak demand

Smart meter data can be used to empirically identify which utility customers contribute the most to peak demand periods. Our work shows that just 10% of households are responsible for nearly 35% of total residential load during peak demand periods and 20% of households account for just over 50% of total residential load. The lowest consuming 50% of households, on the other hand, are responsible for just over 15% of total residential demand. The highest energy use during peak demand is in the Central Valley. Demand response and cooling efficiency programs should be focused in those areas of highest demand. A random sample from the top 10% of peak coincident demand is expected to contribute 3.3x the demand of a sample of similar size drawn from the general population. These basic techniques could be extended to study demand during other time periods when grid resources may be stressed, like morning and evening solar ramps or expected periods of wind variability.

Chapter 3

Recovering and interpreting semi-physical information from residential smart meter data

3.1 Introduction

As smart meter data becomes available for millions of utility customers, uses ranging from more reliable load forecasts to improving the targeting, delivery, and evaluation of energy efficiency and demand response programs are beginning to emerge. Statistical methods of parameter estimation, with regression models prominent among them, are well suited to extracting the information used to support these applications. This paper presents a study of *semi-physical* residential characteristics, defined as the physical processes, building characteristics, and occupant behaviors that shape patterns of energy use, extracted from *daily extracts* of hourly interval data using a regression modeling framework.

Statistical models offer mathematically precise and verifiable properties, but, in non-technical contexts, including the planning, implementation, and evaluation of utility programs, goodness of fit and statistical significance do not necessarily translate into actionable information. This work places a premium on model structures and interpretations informed by the information needs of energy efficiency and demand response programs.

A data set spanning April 2008 to November 2011 from a stratified sample of residential smart meters belonging to approximately 180,000 residential customers of Pacific Gas and Electric (PG&E) provides an opportunity for characterizing different physically plausible regression model formulations. The model fits characterize the patterns of energy use for every residence, thus providing estimates for the distributions of unobserved characteristics of the residences that can be used to support utility program planning and targeting.

This work proceeds as follows: Section 3.2 offers a review of prior statistical modeling techniques applied to building energy use data. Section 3.3 develops semi-physical model features (i.e. the structure of regression model components along with supporting regressor data) based on a building science understanding of the nature of heating and cooling loads, equipment operations, and occupant behavior in homes. These features are then assembled into a selection of regression models designed to emphasize the diversity of potential model configurations without the computational costs of evaluating every permutation. Section 3.4 details the application of these models to validated smart meter data from approximately 160,000 residences, along with the methods and techniques used to characterize their performance. Section 3.5 interprets the technical performance of the models and draws conclusions for modelers. Section 3.6 interprets the modeling results in terms of what they reveal about the residences they characterize and provides estimates of categorically disaggregated energy consumption for all residences in the sample and example applications relevant to assessing program potential and targeting program efforts. Section 5.7 expands the discussion of model results to policy-relevant conclusions and areas of potential future work.

3.2 Prior work

This work presents methods for recovering diagnostic information about individual homes given only hourly interval smart meter readings with timestamps and zip code level location data. These are practical constraints driven by data availability, but they also bolster anonymity[1] and ensure that the methods described are applicable to the millions of homes already equipped with smart meters. Much of the prior work on analysis of whole-building interval meter data has focused on the accuracy of model predictions of consumption for just a few buildings, typically with an emphasis on the impacts of weather. Princeton's PRISM, summarized in [33], uses regression models to fit monthly heating fuel consumption, assuming a temperature-insensitive base level of consumption, α, and a temperature response, β, to outside temperatures linear in heating degree days. Heating degree days (HDDs) are the positive difference between a predetermined threshold temperature, τ, and daily average temperature, T_{out}. It is 0 if the difference is negative and notated as $(\tau - T_{out})_+$. HDD values for a given threshold, τ, are notated as $HDD(\tau)$, so the full regression model used is specified as $E_{monthly} = \alpha + \beta HDD_{monthly}(\tau) + \varepsilon$. The original work correctly points out that τ should be derived from the underlying data because prescribing a value risks biasing other parameter estimates. However, in practice, HDDs are often defined prescriptively using a threshold near 65°F (18°C). The PRISM method can also be applied to monthly cooling energy by using cooling degree days (CDDs), defined as $(T_{out} - \tau)_+$ or $(\tau - T_{out})_-$ for each day. Recently, [96] have introduced a tool that uses monthly utility gas and electric billing data and concepts very similar to PRISM to simultaneously estimate a handful of physical building characteristics capable of diagnosing some categories of control problems.

With the advent of relatively affordable interval metering and computerized controls for larger buildings in the 1990s, models were further refined to accommodate data being collected on hourly and 15-minute timescales. ASHRAE sponsored two energy data prediction "shootouts," [57, 40], using interval meter data from commercial buildings, with the objective of identifying models with the best predictive power. The challenges used data from just a few buildings. The specified performance metrics — cross-validated coefficient of variation and mean bias error, emphasized prediction accuracy.

The winning models, employing variations on neural network techniques, predicted well, but their model parameters, the weights of connections between network nodes, lacked physical significance. Subsequent work, for example [28], suggests that data from other buildings can be better fit using different techniques, albeit with narrow margins of victory.

The class of predictive models that are difficult to interpret in terms of the drivers of energy use can be labeled *black box* models. Their sole objective is prediction accuracy, and the mechanism responsible for their predictions contains little information about the

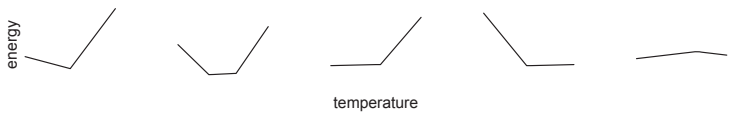
[1] Technically, such data can be re-identified if sufficient information exists in the public domain to correlate patterns of consumption with publicly identified customers, but this is not a concern taken up here.

system being modeled. Models whose parameter fits yield meaningful information about the systems they model are often called *gray box* models. Those whose primary purpose is to deliver meaningful parameter fits are called *inverse models* because they work backward from observations to reconstruct system parameters.

With a focus on physical interpretation, [94] updated the standard PRISM model to allow the use of separate values of HDDs derived using occupied and unoccupied days for 50 commercial buildings. They cautiously interpreted the temperature thresholds and regression coefficients as physically meaningful information and derived estimates of response to outdoor temperature, threshold temperature, and the fixed effects of occupancy. [11] are also interested in physical interpretation. They employed heuristic techniques to separate out base load from variable load and identify heating and cooling temperature slopes and cooling change points for 327 households.

[52] describe the methods used by ASHRAE's Inverse Modeling Toolkit, which is a set of standardized regression tools for fitting interval-meter data. To support hourly timescales, they adopt a change point, rather than a degree-day modeling approach to outdoor temperature response. Their models assume a piecewise linear fit of temperature response, with one or two breakpoints, corresponding to heating and cooling thresholds.

The response to outside temperature therefore has up to three segments. For two segment cases, one of the segments can be fixed to zero thermal response to capture temperatures below the cooling threshold or above the heating threshold without conditioning. Alternately, the temperature response of both segments can be fit by the model to capture both heating and cooling. The three-segment model consists of heating and cooling segments separated by a zero-slope segment for temperatures without conditioning. Table 3.1 provides a schematic visualization of the five major classes of temperature response involving breakpoints. The change points are determined through an exogenous process, prior to model fitting.



Both heating and cooling	Both (separated)	Cooling only	Heating only	Neither

Table 3.1: Schematic diagrams of energy vs. temperature responses for each class of change point response.

Gray box models can be built based on physical heat transfer equations for buildings, often utilizing thermal networks, with equations similar to those used to model the dynamics of RC electrical circuits. For example, models developed by [79] and [70] can be used to estimate energy consumption and building attributes and state, including

internal temperatures, levels of insulation and thermal mass, and HVAC system state, depending on what data is available. Researchers like [93] have also proposed the use of ARMA/Fourier models, calibrated simulation, and other techniques for capturing building dynamics. Recently [7] used model selection techniques to choose the best fitting RC circuit among a selection of options and achieved satisfactory recovery of physically significant building characteristics.

However, thermal network techniques are not directly applicable to the study of buildings with only whole-building meter data available. They tend to require reliable information on physical characteristics of the building as inputs and often rely on sub-metered energy data and data from temperature sensors. They are focused on transient thermal properties of buildings, complete with occupant activity and unseen air flow and other gains, which are notoriously difficult to estimate in practice. Even though they are not directly applicable, the physical concepts and intuition such models develop are highly relevant to the specification and interpretation of less fine-grained models.

Finally, work on non-intrusive load monitoring (NILM) has focused on disaggregating the energy use of specific end uses from whole-building data. [118] reviews the relevant literature on different approaches and [5] make the case that such disaggregation of loads is the "holy grail" of energy efficiency. However, NILM can suffer from errors of kind and magnitude, assigning energy to an appliance category not present in the home or missing consumption of one that is. The accuracy of NILM methods is strongly dependent on the sampling rate of the underlying signal. Algorithms requiring data sampled at 10kHz or more, for example [35], perform quite well. Unfortunately, these sampling rates and associated computational power are incompatible with existing metering infrastructure[2]. Algorithms that work with data sampled at hourly intervals, for example [54], can only be expected to achieve classification accuracies of 50-60% of energy correctly classified. Even with these errors, the information contained in NILM results should still be somewhat useful in estimating stock characteristics of specific appliance categories and identifying potential target homes for efficiency and demand response programs. However, they cannot provide enough certainly about end uses to act with confidence. The quest for the holy grail of accurate disaggregation continues.

3.2.1 Our contribution

The goal of residential efficiency programs is to replace energy using devices or change how they are operated so the services desired by residents can be provided with less waste. The goal of demand response programs is to establish control over energy using devices so that their operations can be altered to reduce demand during time limited periods of grid stress while maintaining a minimum standard of service. The design, implementation and evaluation of utility programs therefore require information about the ownership of

[2]Because NILM is supposed to be an alternative to installing metering hardware, it is unsatisfying to contemplate hardware upgrades to support it.

major energy using devices and their patterns of use, as dictated by the preferences of their owners.

This work develops and applies regression models of smart meter data whose coefficients recover information directly responsive to utility program information needs. The models have been designed to emphasize physical interpretation of results, the ability to run against large sample sizes (approximately 160,000 homes with at least 180 days of hourly data in this case), and to allow assessment of relative and absolute performance of different model specifications. Models cover several formulations of thermal response and change points, differentiate days of the week, and account for length of daylight and the possibility of thermal lags. Household electricity consumption data provided by smart meter infrastructure provides enough samples to develop detailed distributions of model and parameter fits across the households sampled from the PG&E service territory. These empirical distributions provide information about the nature of the building stock as well as the relative strengths and weaknesses of the underlying models. Subject to the limitations of the accuracy of the modeling techniques, parameter estimates provide information about individual homes sufficient to inform potential estimates and targeting for energy efficiency and demand response efforts.

3.3 Defining physically significant regressors

This section provides a summary of the dynamics expected to influence daily patterns of electricity use and describes the specification of semi-physical regressors and the models composed of those regressors that are used in this study. Because each residence's pattern of electricity demand is determined by a unique combination of physical characteristics, equipment, controls, and occupant behaviors, specific combinations of regressors will perform better or worse in explaining or predicting meter data from each home. For example, a regressor based on outside temperature will explain a significant amount of variation in a residence with air conditioning and much less in a residence without air conditioning or electric heat.

3.3.1 Hourly vs. daily dynamics

When given a data set with fine time resolution, it is tempting to dive in and build models that explain every wiggle in demand. This can be a difficult process because so much home energy use is subject to the unobserved state and physical characteristics of the buildings and the whims of the occupants. It is stochastic in both time and magnitude.

Internal temperatures are driven, in part, by changing outside conditions, mediated by insulation and thermal mass — processes that tend to have time constants of hours or days. Transient states will tend to confound models with shorter time frames. As the ASHRAE Handbook of Fundamentals puts it, *"Variations in the characteristics of residences can lead to surprisingly complex load calculations. Time-varying heat flows*

combine to produce a time-varying load. The relative magnitude and pattern of the heat flows depends on the building characteristics and exposure, resulting in a building-specific load profile."

A model based on total electric energy consumed each day, in kWh, averages out transients that take place over shorter time scales, including occupant schedules, thermal lags, and ad-hoc consumption during the course of each day, yet they preserve the signatures of thermal response and occupant activity at the daily level. A model focused on a home's energy response to outside conditions at finer time scales must expend significant effort performing state estimation, often without the benefit of ground truth data.

Impact of time averaging on electricity consumption data

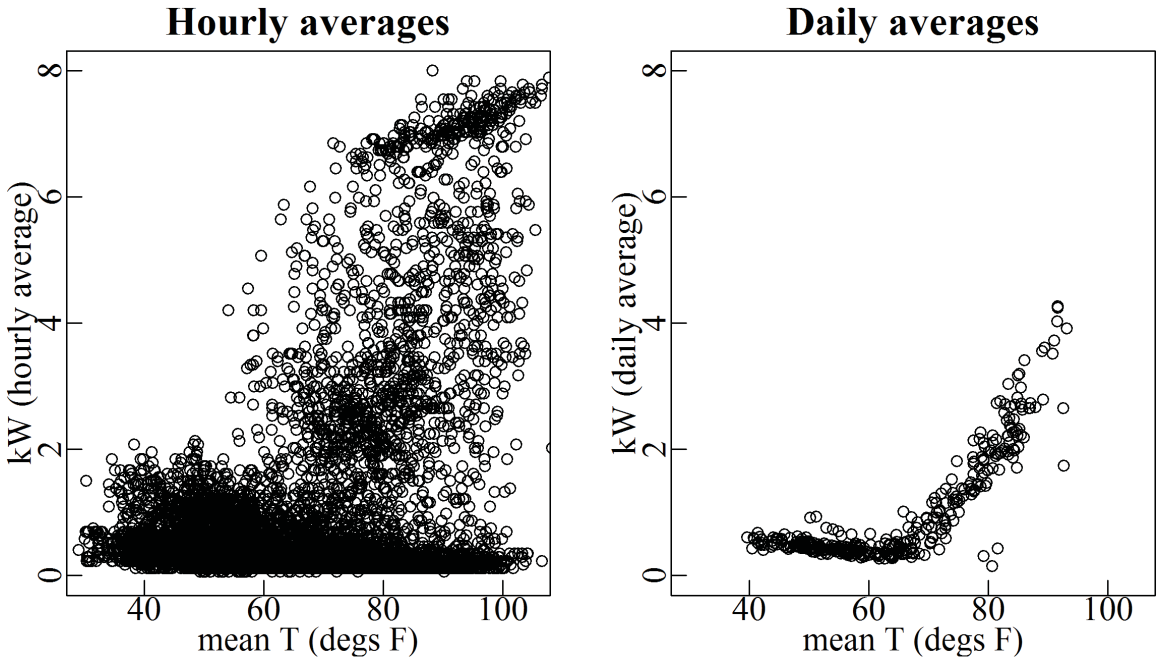


Figure 3.1: Scatter plots of average power demand vs. average temperature for hourly (left) and daily (right) time intervals for the same residence.

Figure 3.1 on page 53 provides insight into the challenges of modeling hourly electricity demand. Both scatter plots draw upon the same data from a residence located in the hot climate of Bakersfield, CA (California's Zone 13). The hourly data on the left suggests highly variable consumption, ranging from a consistent and temperature-insensitive minimum well under 1 kW to temperature-sensitive maximum values that increases dramatically at higher temperatures to an upper limit of between 7 and 8 kW. In between these extremes are many interstitial observations. This pattern is consistent with consumption influenced by occupancy and scheduling, mediated by the mass of the walls

and interior furnishings, and observations that, at hourly intervals, span different portion of the duty cycle of the air conditioning unit. The daily averages of the data, on the other hand, are quite well clustered around two linear segments with differing slopes, intersecting at a break point just under 70°F (21°C). The lower temperature segment rises slightly with cooler temperatures, suggesting a modest temperature sensitivity, possibly associated with the blower fan of a natural gas furnace. The upper segment increases with increasing temperatures, suggesting growing demand for air conditioning as temperatures rise. This household is a good match for the "cooling only" class of change point response. In this case, the highest 24-hour mean temperatures exceed 90°F (32°C), virtually guaranteeing a need for aggressive cooling, which pushes average daily household demand to approximately 4 kW, or nearly 100 kWh/day.

Transient states, thermal or otherwise, create significant amounts of variability and temporal correlation in observations sampled faster than the time constants that characterize the transients, and thus confound stateless models. In the case of regression models, it is theoretically possible to observe operational and thermal states and include them as regressors[3]. However, practical limitations of time and resources often limit the availability of such data, as is the case with smart meter data.

Models based on daily averages of outdoor conditions and daily totals of energy use avoid many of the problems created by transient states, while still allowing for a detailed study of thermal and temporal patterns of consumption. They also reduce the amount of data fed into the regressions by a factor of 24, greatly increasing the number of modeling runs that can be performed in a fixed amount of time. For these reasons, the models used for this study are based on daily observations derived from aggregated hourly data.

[3]For example, we would want to know whether windows are open, how much solar radiation is hitting the roof or streaming through the windows, which appliances are on, indoor temperatures, thermostat settings, internal wall temperatures, ground temperatures, etc.

Typical daily kWh vs. daily mean outside temperature scatter plots for 9 residences

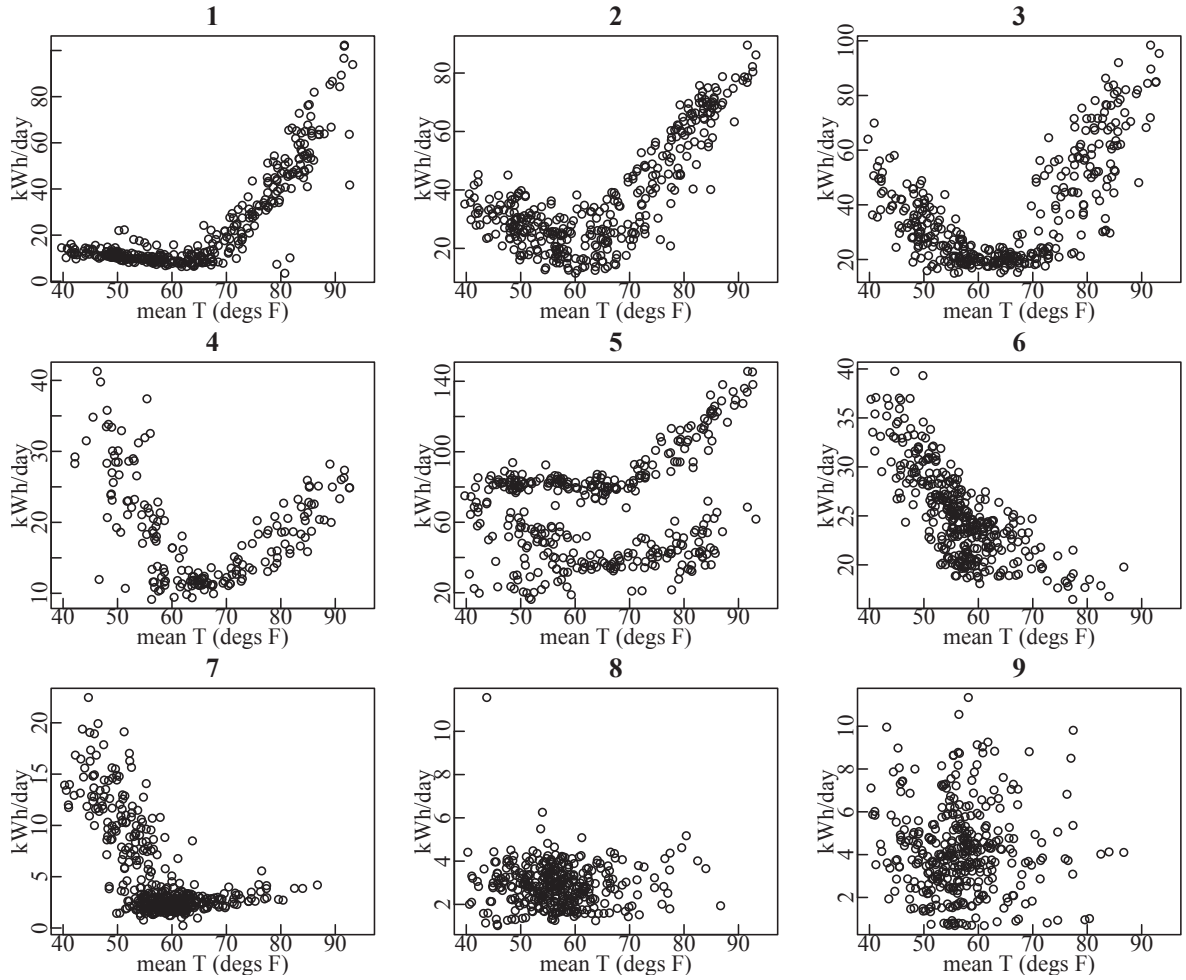


Figure 3.2: Examples of daily energy scatter plotted against daily mean temperature, illustrating the presence or absence of change points for heating and cooling and other patterns of consumption in a range of buildings.

Even though the use of daily data clears up several modeling concerns and results in simpler model specifications, there is still a lot of variation between residences. Figure 3.2 on page 55 provides a look at the daily energy vs. temperature scatter plots for nine residences selected to highlight the diversity of thermal responses found in the data. First note that the y-axes are scaled to accommodate a substantial range in daily energy use. Sub-plots from residences 1-5 exhibit high temperature cooling. 2-4, 6, and 7 exhibit evidence of electric components of heating. 8 and 9 show no relationship between temperature and energy consumption. 3 and possibly 4 are more U-shaped than V-shaped,

evidence of a separation between their heating and cooling change points. 5 seems to have some fixed amount of energy consumption added to some days but not others, which produces two separate clusters of points, possibly driven by differences between weekends and weekdays, with very similar thermal characteristics, but separated by a fixed amount of energy. The diversity revealed by these plots strongly suggests that no single model will perform well explaining the patterns of all homes. This study utilizes a framework that evaluates several model formulations against each home.

3.3.2 Categories of semi-physical information

This section provides descriptions of the categories of semi-physical information our models are designed to recover. The categories are based on the information requirements of utility programs but also informed by the limitations of what can be learned from relatively coarse meter data.

Table 3.2: Characteristics of household energy use that help provide a physical explanation of daily totals of home energy usage.

Name	Description
Thermal response	Heating and cooling electricity usage, when present, will be roughly proportional to the rate of heat losses or gains, which is proportional to the difference between indoor and outdoor temperatures. All else being equal, cooling energy will increase as average daily outside temperatures increase. Depending on the type of heating fuel, electricity consumption can also be proportional to heating demand (increasing as temperatures decrease), either capturing the full demand in the form of resistance heat or heat pumps, or just a fraction of the demand via the supporting fans and pumps of direct combustion heating systems.

Change points	Because they are driven by comfort, heating and cooling are only necessary when a home would otherwise be too cold or hot, respectively. As a result, such loads tend to be controlled by thermostats or operated manually in a manner consistent with heating and cooling set points. While set points maintain indoor conditions, indoor conditions are strongly influenced by outdoor ones (mediated by shell characteristics and internal gains). Thus, every set point has a corresponding balance point in outdoor temperature where the conditioning system shifts from off to on. These outdoor temperature balance points become the change points in models of piecewise temperature response.
Thermal lags	The flow of thermal energy back and forth from outside to inside a home is mediated by the thermal mass of the building materials in walls, floors, ceilings, basements, and attics. Thermal mass has the effect of dampening and delaying the transmission of heat and cold from the outside. Depending on construction details, the delay can span from one or two hours to more than a day. However, direct openings and highly transmissive materials can circumvent much of this effect. A much longer seasonal lag characterizes the thermal exchange between a home and the ground beneath it.
Lighting	While lighting has many uses, much of the lighting energy in a typical home is used for illumination after dark. It stands to reason that lighting should be used for more hours of the day during the long nights of winter than the shorter nights of summer and this effect should be discernible in daily patterns of energy use.
Minimum loads	Some loads are never turned off. These include refrigerators, some kinds of lighting, standby (or active) power for various electronics, transformers in power adapters, clocks, computer networking equipment, etc. Because they are always on, such loads can contribute a significant fraction of total energy consumption — even if they are rarely interacted with directly. Utility concerns about system capacity are driven by instantaneous power demand, while concerns about pollution, fuel usage, and costs tend to be driven by total energy, which is, in turn, often dominated by these steady, long-term energy loads.

Occupancy	Many large loads in a home are under direct occupant control, with the timing and magnitude of use directly impacted by the decisions of the members of the household. Hair dryers, blenders, laundry and dishwashers, televisions, and many other loads may be roughly predictable in general terms, but usually with significant uncertainty in both magnitude and timing. Thus, human behavior can be expected to contribute significant variability to patterns of energy demand, with some households displaying more regular patterns of usage than others. Occupant preferences also determine which equipment is present in a home and, in the case of heating and cooling, what set points and schedules are used. Resolving occupant-driven loads fully would require extremely detailed knowledge of occupant schedules, behaviors, vacations, etc. As a substitute, normative assumptions about patterns of occupancy can be built into model formulations based on timing of use within days, across days of the week, or seasonally.
Day of week	One obvious way occupancy varies is by the day of the week. For example, many households follow different schedules on weekdays and weekends. However, the nature of those schedules varies by household. Some people spend more time at home on weekends, others spend their weekends away from home. Further complicating daily patterns of energy use, many people work at home, don't work, or work non-traditional hours and days. At the daily level of data aggregation, outliers are often generated by unusual occupancy. Holidays are one form of more or less predictable changes in occupancy, but many outliers are not responsive to any known external condition.

3.3.3 Candidate regressors

The challenge of formulating effective regression model structures with semi-physical parameter fits centers on the development of regressors that reflect the physical knowledge outlined above. In this section we present detailed specifications for the regressors used by the regression models in this study, followed by the model formulations themselves.

3.3.3.1 Notation

The models are based on the assumption that each observation of daily energy consumption can be modeled as the weighted sum of multiple regressors (assume there are k of them) observed on the same day, i:

$$kWh_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{i,j} + \varepsilon_i \tag{3.1}$$

To support the accurate specification of the regressors, a bit of additional notation is required.

1. It is common in regressions to use indicator variables as regressors that are either 0 or 1 depending on specific conditions. The corresponding coefficients have the same units as the explained variable (kWh/day in this case) and can be interpreted as the effect of the presence of the condition that causes the regressor to be equal to 1. Examples include coefficients for the days of the week or other known occupancy states. Here, the notation for indicator variables makes use of delta functions that evaluate to 1 for all observations where a given expression is true and 0 otherwise. For example a variable called "day" that assigns days of the week using numbers from 0 to 6 can be broken into seven columns of indicator variables whose values are 1 for every observation of "day" matching their column number and 0 otherwise. In more formal notation, the 7 columns of data look like:

$$\sum_{d=0}^{6} \delta_{day=d}$$

2. To implement piecewise linear change point responses to outside temperature, temperature observations must be manipulated into one column of data per segment. We introduce notation to handle this manipulation. For the vector of temperature observations, T, and a change point of 65°F (18°C), we define $(T-65)_+ \equiv max(T-65, 0)$ to produce a new vector composed of just the positive values of the difference between T and 65°F. For the regressors described below, some of which describe a piecewise linear temperature response, a more flexible notation is required. We define $T^*_{T_1,T_2} \equiv min((T-T_1)_+, T_2-T_1)$, where $T_1 < T_2$. Therefore when T_1 and T_2 are finite, $0 \geq T^* \geq T_2-T_1$. The special case of $T^*T_{1,\infty} = min((T-T_1)_+, \infty) = (T-T_1)_+$. The special case of $T^*_{-\infty,T_2}$, which technically evaluates to ∞, is defined to equal $(T_2 - T)_+$ instead. This fixes the zero point of thermal response to the lowest (or only) break point.

3.3.3.2 Regressor definitions

Table 3.3: Names and definitions for the candidate semi-physical regressors used to generate the models evaluated in this work. Elsewhere in this text, the names of regressors are used as shorthand for the underlying data, model structure, and resulting regression coefficient.

Regressor name	**Description**	**Regression formula notation**
Indicator variables	**0 or 1 indicate the presence or absence of specific conditions**	
`DOW`	Day of week – one indicator variable column for each day of the week	$\sum_{d=0}^{6} \beta_{DOW_d} \delta_{day=d}$ for $day \in \{0..6\}$, with 0=Sunday through 6 = Saturday for all date observations. Note that $\sum_{d=0}^{6} \delta_{day=d} = 1$ for any day of the week, so the Sunday column is typically dropped to avoid co-linearity with the constant term in the regression.
`WKND`	Weekend vs weekday – a single indicator variable with value 1 on weekend days and 0 on weekdays	$\beta_{wknd} \delta_{day \in \{0..6\}}$ for day defined as numbers 0 through 6, with 0=Sunday through 6 = Saturday for all date observations.
Standard variables	**Smoothly varying variables**	
`DL (day length)`	Number of hours each day is longer than the shortest day of the year (computed as the length of daylight for a given day minus the length of daylight on the winter solstice)	$\beta_{DL} DL$, where DL is the number of hours the day is longer than the winter solstice.

`CDH (cooling degree hours)`	The sum across each 24-hour day, weighted by the positive difference between 65°F (18.3°C) and hourly temperature readings.	$\beta_{CDH}CDH$ where $CDH_i = \sum_{HOD=0}^{23}(T_{i,HOD} - 65)_+$ for hourly observations of temperatures, T grouped by day i.
`tout`	The mean hourly outside temperature for each day. A variation on t_{out}.	$\beta_{tout}tout$ where $tout_i = \sum_{HOD=0}^{23} T_{i,HOD}/24$ for hourly observations of temperatures, T, grouped by day i
`tout.max`	The maximum hourly outside temperature for each day	$\beta_{tout.max}tout.max$ where $tout.max = max(T_{i,HOD\in 0..23})$ for hourly observations of temperatures,T, grouped by day,i
`tout.min`	The minimum hourly outside temperature for each day	$\beta_{tout.min}tout.min$ where $tout.min = min(T_{i,HOD\in 0..23})$ for hourly observations of temperatures,T, grouped by day,i
`CP65`	Single column of values corresponding to the number of degrees the mean daily temperature is above 65°F (18.3°C), fixed to zero if the value would be negative.	$\beta_{65+}(tout - 65)_+$
`toutCP`	Two columns of temperature data corresponding to a piecewise temperature response broken at a specified temperature (incremented in 1°F steps). The change point, CP, is set to the temperature that minimizes the Sum of Squared Residuals for the data at hand, as determined by a grid search of whole number temperatures.	$\sum_{i=1}^{n} \beta_{c_i,c_{i+1}} tout^*_{c_i,c_{i+1}}$ for $c = \{-\infty, CP, \infty\}, n = 2$ The lower column is $(CP - tout)_+$ and the upper column is $(tout - CP)_+$.

`toutCP65`	Two columns of temperature data corresponding to a piecewise temperature response broken at 18.3°C (65°F), which is often used as a presumed break point absent evidence to the contrary.	$\sum_{i=1}^{n} \beta_{c_i,c_{i+1}} tout^*_{c_i,c_{i+1}}$ for $c = \{-\infty, 65, \infty\}, n = 2$
`tout2CP`	Three columns of temperature data corresponding to a piecewise temperature response broken at two separate temperatures (incremented in 1°F steps) that minimize the Sum of Squared Residuals for the data at hand.	$\sum_{i=1}^{n} \beta_{c_i,c_{i+1}} tout^*_{c_i,c_{i+1}}$ for $c = \{-\infty, CP_1, CP_2, \infty\}, n = 3$ For two given change points, CP_1 & CP_2, the lower column is $(CP_1 - tout)_+$, the middle column is $min(CP_2 - CP_1, (tout - CP_2)_+)$, and the upper is $(tout - CP_2)_+$
`toutFCP`	Four columns of temperature data corresponding to a piecewise temperature response broken at 55,65,75°F (13,18,24°C). FCP stands for fixed change points to distinguish this one-size fits all regression from those that separately estimate the change points as exogenous model parameters.	$\sum_{i=1}^{n} \beta_{c_i,c_{i+1}} tout^*_{c_i,c_{i+1}}$ for $c = \{-\infty, 55, 65, 75, \infty\}, n = 4$
`L1 (lag 1 of CP65)`	`toutCP65` lagged by one day (i.e. the previous day's mean temperature with a break at 65°F (18.3°C)	$\beta_{L1} L1_i$ where $L1_i = (tout_{i-1} - 65)_+$

For convenience, regressors and models will be discussed by name for the remainder of this work. For example, the model

$$kWh = \beta_0 + \sum_{d=1}^{6} \beta_{DOW_d} \delta_{day=d} + \beta_{tout} tout + \beta_{DL} DL + \varepsilon \tag{3.2}$$

has been assigned the name `DOW+tout+DL` because, drawing upon the definitions above, it is composed of the terms `DOW`, `tout`, and `DL`. For the rest of this paper, the model names will be used as a shorthand for these more formal model definitions. In the context of model results, the names will be used to refer to the regression coefficients. In the context

of model descriptions, the names refer to the model structure and related underlying data. For example, the model defined by equation 3.3 is named `DOW+toutCP+DL`.

$$kWh = \beta_0 + \sum_{d=1}^{6} \beta_{DOW_d} \delta_{day=d} + \beta_{tout\_}(CP - tout)_+ + \beta_{tout_+}(tout - CP)_+ + \beta_{DL} DL + \varepsilon \qquad (3.3)$$

3.3.4 Model specifications

The models evaluated in this work are composed of different permutations of the regressors described in the previous section. They are designed to span a large range of logical patterns of residential energy use, to ensure that many models have nested relationships, and to ensure that the impacts of regressors of particular interest can be studied in isolation or in the context of a related group of regressors. Another significant concern was limiting the number and complexity of models to allow for a manageable run time for all the models across all 160,000 residences[4]. Table 3.4 on page 63 provides a summary of the names and composition of the regression models used for this study.

model name	WKND	DOW	day. length	tout	tout.min	tout.max	CDH	tout65	L1	toutCP65	toutCP	tout2CP
tout				x								
WKND	x											
DOW		x										
DOW+tout		x		x								
DOW+tout+DL		x	x	x								
DOW+tout.min+DL		x	x		x							
DOW+tout.max+DL		x	x			x						
DOW+DD+DL		x	x				x					
DOW+tout+tout65+DL		x	x	x				x				
DOW+tout+DL+L1		x	x	x					x			
DOW+toutCP65+DL		x	x							x		
DOW+toutCP+DL		x	x								x	
DOW+toutCP+DL+L1		x	x						x		x	
DOW+tout2CP+DL+L1		x	x						x			x
DOW+toutFCP+DL+L1		x	x						x			

Table 3.4: Model names and their regressors for the candidate models used for this study.

[4]The R code used to perform these analyses was written with performance in mind but is in no sense fully optimized. A complete run of all models using four 2 GHz processor cores takes approximately 1 week to execute. However, cross validation is responsible for most of this run time.

3.4 Running the models

The data access, cleansing, validation, and statistical modeling for this work were all performed in the open source statistical computing platform R version 2.15.2. See [92] for details on the platform. Much of the data visualization was performed using the powerful and elegant R plotting module `ggplot2`, documented in [117]. The specified models were run, using R's `lm` function, against daily averages for a sample of hourly interval meter data from Pacific Gas and Electric (PG&E). The raw sample consisted of a sample of 180,000 residences from 588 five-digit zip codes of the service territory. The data covers April 2008 to the end of November 2011. The sample was stratified to ensure well specified numbers of accounts associated with single family residences, landlords, and renters. Validation rules mandating at least 180 days worth of data, no more than 15% of readings equal to 0, and mean power demand greater than 110W narrowed the results to approximately 160,000. For applications requiring estimates of annual consumption, the sample is further reduced to 128,000.

According to PG&E's 2009 SEC 10K filing, [90], the company served nearly 4.5M residential customers, or about 28 times the number of households in our sample. In 2009, all of PG&E's residential customers consumed approximated 31,000 GWh of electricity, which is 36% of PG&E's total for the year. This is an average of nearly 6,900 kWh per household. The average for the validated data sample used for this study is 6,960 kWh/year, perhaps slightly higher due to the validation rule that eliminates extremely low usage.

Day of week and day length regressors were derived from data time stamps. Hourly temperature data was collected from online weather data aggregator Weather Underground, with simultaneous values from 3-5 nearby weather stations averaged, dropping missing values, to produce a separate time series for every zip code.

Significant programming effort went into data access, pre-processing, storage, post-processing, analysis, and visualization of the data and modeling results. The following pseudo code listing captures the basic structure of the algorithm that ran the models.

```
modelSpecifications = <definitions of models to be run>
zipCodes = <all zip codes in data set>
modelRuns = resultSet()          # Data structure to hold the model runs
for zip in zipCodes {            # About 580 zip codes
  zipMeters = metersIn(zip)      # All the data associated with the zip code
  weather = weatherDataFor(zip)  # Weather spanning all the meter reading dates
  for meter in zipMeters {
    if(validated(meter)) {       # Ignore invalid data
      basicStats = basicAnalysis(meter.data)
      modelResults = runModels(meter.data,weather,modelSpecifications)
      modelRuns.append(meter.id,basicStats,modelResults)
    }
  }
}
<save modelRuns to disk>
# We now have outcomes from the specified models applied to
# data from every available meter.
```

3.4.1 Defining and measuring model performance

This work makes extensive use of regression models to fit parameters that explain time series residential energy data. Our primary concern is that the models are formulated to provide physically meaningful information about the patterns of energy use in the metered residences. The assessment of those model formulations requires a treatment of standard regression modeling concerns, however, the derivation of meaningful information does not necessarily require good model fits. Because our models can be interpreted as normative descriptions of patterns of residential equipment use and occupancy, weaker fits provide evidence for patterns that differ from those normative assumptions. In other words, there is information in bad fits as well as good ones.

Past regression modeling efforts have typically focused on one or just a handful of buildings, allowing models to be hand tuned with careful attention paid to metrics of model performance and the plausibility of parameter fit values. Here we have well over 160,000 runs of each model specification, so model comparisons must be algorithmic, with overall performance judged using aggregated models. Individual outcomes can be evaluated within the context of a large population of results that reveal the underlying character of each model specification and suggest when specific formulations can be expected to perform well.

The usual model performance concerns, including the potential for mis-specified models and the presence of serial correlation or non-normal error distributions, are still operative. Of these concerns, we are most interested in assessing model specifications. The following enumeration describes the three broad categories of potential errors in the context of this work.

1. **Mis-specified models** — As the term implies, mis-specified models do not achieve accurate coverage of the factors that govern the data-generating process they attempt to explain. This can undermine model fits through unreliable parameter and error estimates, and symptoms of serially correlated and non-normally distributed errors. We are particularly interested in the potential for omitted variables. Our observations are limited to daily summaries of hourly energy use and regressors that can be derived from the associated timestamps and location at the zip code level. We have no direct information about occupants, building construction, or household appliances and controls.

2. **Serial correlation of errors** — Correlated errors can be caused by mis-specification of a model. In our case, failing to account for changes in how each home is operated over time, for example due to seasonal occupancy or the birth of a child, is a significant model specification concern. However, even with well specified models, serial correlations can arise from correlations in the underlying data, which are endemic to time series data. We expect correlations in weather data, occupancy, and energy use from day to day (and certainly hour to hour). Models built to explain serially correlated data will tend to be over confident in their metrics of model fit and the

standard errors of parameter estimates. However, serially correlated regressors can also lead to over estimation of standard errors.

3. **Non-normal distribution of errors** — Normality of regressors and errors is a basic assumption of standard least squares regression. Non-normality can be caused by mis-specified models, outlier data, or distributions of observed variables that are themselves non-normal. We know from experience that regressions of energy use data tend to have fatter tails in their error distributions than are found in normal distributions. This suggests that inferences that rely on the normality of errors, for example the p-values of the coefficient fits, will underestimate the uncertainty of the fit.

3.4.1.1 Metrics of performance

This section provides details on the calculation of the metrics used to assess model performance. Some common metrics have unambiguous definitions, but most can be found in more than one form in the literature. The following definitions and discussion of the metrics used provide the formulations used as each metrics was gathered for every model run. First, some notation. For every observation y_t at time t, the model estimate is notated as $\hat{y}_t$ the error at time t is defined as: $e_t = y_t - \hat{y}_t$. The mean of all the observations is $\bar{y}$. The number of observations is n and the number of model parameters is k.'

Table 3.5: Names, formulas, and descriptions of metrics of model fit used elsewhere in this text.

Metric	Formula	Description
RMSE	$\left(\frac{1}{n}\sum_{t=1}^{n} e_t^2\right)^{\frac{1}{2}}$	The root mean squared error is our primary model error metric. It has units of kWh/day and is therefore appropriate for assessing uncertainties for questions relating to power or energy. Smaller values are better.

cvRMSE	multiple out of sample applications of RMSE	Because we are concerned about the possibility of our models over-fitting the data or being too impacted by outliers, we calculate a cross validated version of RMSE (called cvRMSE in this text). The cross validation is intended to penalize models that over-fit data and provide a more accurate metric of how well a model predicts energy use, as opposed to explaining the observed data. Cross validation randomly assigns observations to five equally sized groups. Five separate regressions are performed, each one withholding the data from one group, and the RMSE is calculated using model predictions for the withheld data. The five resulting prediction errors are averaged to provide the cvRMSE for the model. Because there is an element of chance introduced by the random assignment of groups, this whole process is repeated four times, with the average of these runs reported as the cvRMSE. As a measure of prediction error, cvRMSE is always larger than RMSE.
cvMAPE	$\frac{1}{n}\sum_{t=1}^{n}\frac{abs\left(y_t-\hat{y}_t\right)}{y_t}$	Mean absolute percentage error is the average of the percentage each model prediction is different from the observed values. Here it is cross validated using a process similar to cvRMSE. Smaller values are better.
R^2	$1-\frac{\sum_{t=1}^{n}e_t^2}{\sum_{t=1}^{n}\left(y_t-\bar{y}_t\right)^2}$	For least squares regression, this is the ratio of explained variance to total variance, varying from 0 (the model explains none of the variance) to 1 (the model explains all of the variance). Unlike RMSE, R^2 is unit-less and normalized by the magnitude of the explained data. R^2 is often used as a metric of model fit, but it tends to reward serial correlation, extra parameters, and over-fit models. Larger values are better.
adjR^2	$R^2-\frac{k\left(1-R^2\right)}{(n-k-1)}$	Adjusted R^2 adjusts R^2 with a penalty for model complexity. Note that for large n (i.e. small k/(n-k)), adjusted R^2 converges to R^2 . In practice, for our samples with hundreds of daily observations, the adjusted R^2 does not alter R^2 by very much. Larger values are better.

AIC	$-2ln(L) + 2k$ or $n\,ln(\frac{1}{n}\sum_{t=1}^{n} e_t^2) + 2k$	The Akaike Information Criterion is a metric used to compare the relative quality of statistical models. It does not require the models to be nested. It is based on the log of the statistical likelihood that the observed data was generated by the model proposed, altered by a term that penalizes model complexity. Because it is a relative metric, its absolute value is not meaningful and definitions of AIC can differ by a constant. For linear regression models, this calculation simplifies to the 2nd formula. Smaller values are better.
kurtosis	$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3$	Excess kurtosis, defined as $\frac{\mu_4}{\sigma^4} - 3$ where μ_4 is the 4th moment about the mean of the values and σ is the standard deviation is a measure of how rounded or peaked an empirical distribution is. The 3 is subtracted because the kurtosis of a normal distribution is 3. Thus, a value of 0 corresponds to normally distributed errors, positive values correspond to sharper peaks and fatter tails, negative values correspond to a rounded peak and thinner tails. Smaller absolute values are better.
DW	$\frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$	The Durbin-Watson test statistic quantifies the degree of serial correlation in neighboring regression errors. It ranges from 0 (perfect correlation) to 4 (perfect anti-correlation), with 2 corresponding to no correlation. Values closer to 2 are better.
lag1cor	$Cor(e_t, e_{t-1})$	The lag 1 correlation between sequential errors is a simple metric of serial correlation. Correlation values of 1 indicate perfect correlation, 0, no correlation and -1, perfect anti-correlation. Smaller values are better.

3.5 Results: Model performance

This section applies the results of the model runs to the question of how well the models performed against a large set of empirical data. These suggest which regressors are useful in explaining household energy use and which households are best explained by which model formulations and have implications for the design of models that explain or predict household energy. The lessons learned are applicable to the common practice of weather correcting energy data, the evaluation of energy efficiency program impacts, and the estimation of heating and cooling demand.

3.5.1 Distributions of model fits

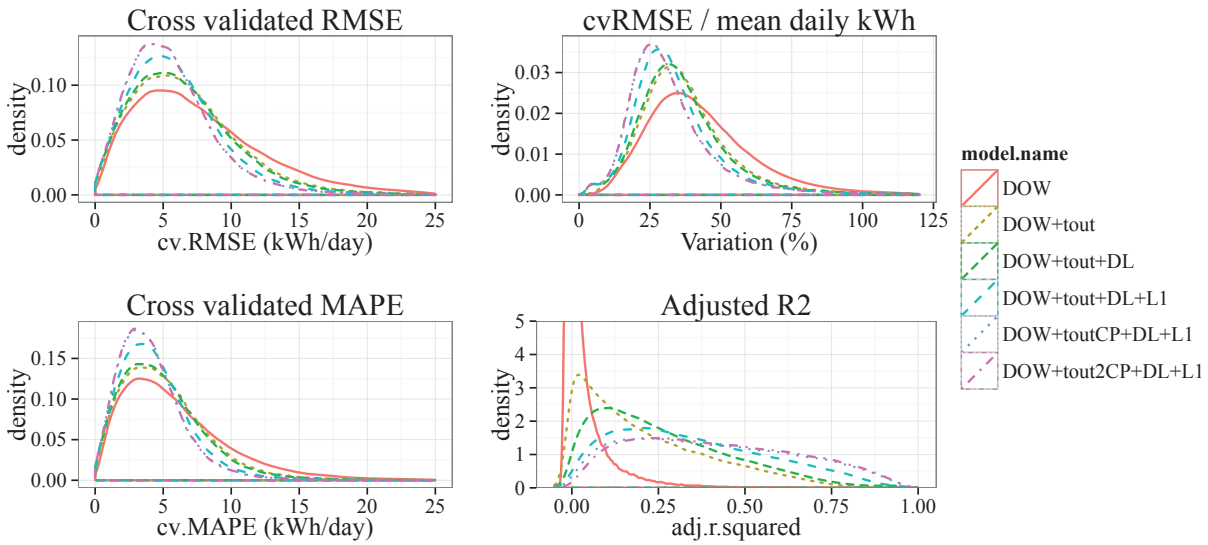


Figure 3.3: Comparison of distributions of model fit metrics for the sampled population of residences across several model formulations.

Figure 3.3 on page 69 shows the cvRMSE (lower is better), normalized cvRMSE (lower is better), cvMAPE (lower is better) and adjusted R^2(closer to 1 is better) for several models selected to span the range of model specifications. The two models with change points, `DOW+toutCP+DL+L1` and `DOW+tout2CP+DL+L1`, have the best overall performance, followed by `DOW+tout+DL+L1`. Recall that the `L1` term is the prior day's cooling degree days with a threshold of 65°F. This is a form of change point as well.

Because heating and cooling systems operate relative to specific set points, it should come as no surprise that change-point models perform well. Another pair, `DOW+tout` and `DOW+tout+DL`, take temperature into account, but without any change points. They both perform far better than `DOW` alone. It is notable that the distributions of each performance metric have pronounced tails in the direction of lower performance. This suggests that a significant number of households are poorly fit regardless of which combinations of regressors are used. Spot checking of data reveals that households without systematic thermal responses, whose variability is dominated by the ad-hoc behaviors of their occupants, and those with unexplained changes in patterns of energy use within their data tend to populate these tails.

However, there are a significant number of households with very good fits as well. These households operate in a manner similar to what the model definitions assume. This means that they tend to have predictable patterns of energy use as a function of day of week, and if they have temperature responses, those responses are also predictable. In

homes with very large cooling loads, the predictable behavior of the air conditioning can dominate all other sources of variance in the home.

Focusing on the adjusted R^2 metric, we can see that the two best model fits, which are nearly identical, have a very broad and flat distribution of values. This indicates fairly uniform probabilities of coming across households with good or bad fits. However, in this case, we can improve our understanding of the causes of better and worse fits with a different view of the data. Figure 3.4 on page 70 takes the adjusted R^2 distribution for the `DOW+toutCP+DL+L1` model and breaks it into one distribution for each of California's climate zones found in PG&E's territory. This figure strongly suggests that the single-change-point model performs best in climates 11-13. These happen to be the hottest climates, so it is reasonable to expect that air conditioning will operate throughout much of the year and dominate total energy consumption and patterns of variation. This strongly suggests that homes with large conditioning loads are more predictable.

Adjusted R^2 for the `DOW+toutCP+DL+L1` model across climate zones

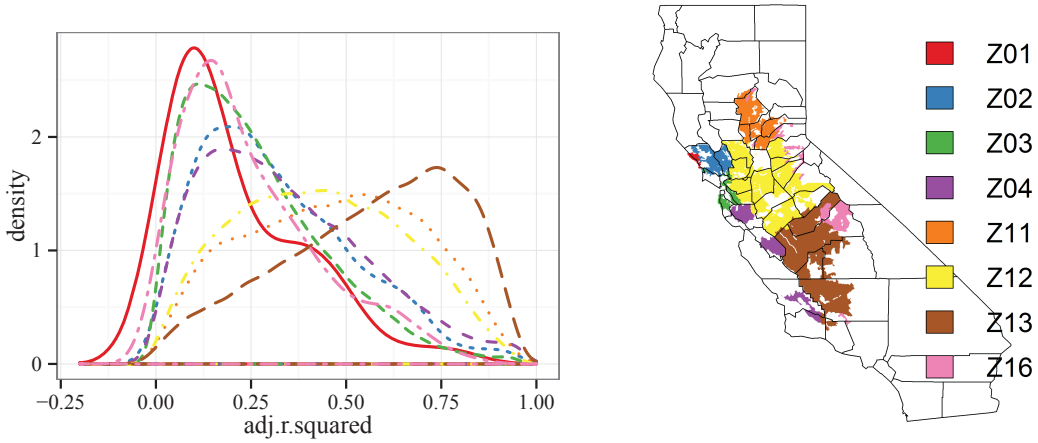


Figure 3.4: The adjusted R^2 performance of `DOW+toutCP+DL+L1`, across the California's climate zones that intersect with PG&E's territory. Zones 1-4 are cool and coastal — they run from north to south consecutively. Zone 16 covers the mountainous regions of the state. Zones 11-13 cover the Central Valley from north to south, with associated increases in temperatures.

3.5.2 Model performance across households

The distributions of metrics above are useful for getting a feel for overall goodness of fit from each model, but they cannot illustrate where each household falls within the distributions. When the fit for a home is greatly improved by the addition of a specific regressor, that reveals something important about how that home consumes energy. This section elaborates on which models fit which households according to a variety of metrics

of model fit.

	R^2	adjR^2	RMSE	cvRMSE	cvMAPE	lag1cor	DW	kurtosis
tout	0.19	0.18	6.94	6.97	5.27	0.52	1.05	3.04
WKND	0.02	0.02	7.94	7.97	6.01	0.59	1.20	3.43
DOW	0.04	0.02	7.93	8.04	6.05	0.61	1.22	3.42
DOW+tout	0.23	0.21	6.84	6.96	5.24	0.53	1.07	3.09
DOW+tout+DL	0.29	0.27	6.61	6.73	5.03	0.50	1.01	3.30
DOW+tout+DL+L1	0.37	0.35	6.07	6.21	4.53	0.41	0.84	3.49
DOW+tout.min+DL	0.26	0.25	6.75	6.88	5.15	0.50	1.01	3.35
DOW+tout.max+DL	0.27	0.26	6.69	6.82	5.10	0.50	1.02	3.32
DOW+CDH+DL	0.35	0.33	6.16	6.27	4.61	0.46	0.93	3.38
DOW+tout+DL+vac	0.29	0.27	6.59	6.74	5.04	0.49	1.00	3.20
DOW+toutCP+DL	0.42	0.40	5.74	5.87	4.26	0.39	0.80	3.65
DOW+toutCP+DL+L1	0.42	0.41	5.69	5.83	4.23	0.39	0.79	3.60
DOW+toutFCP+DL+L1	0.43	0.41	5.67	5.94	4.25	0.38	0.79	3.59
DOW+tout2CP+DL+L1	0.44	0.42	5.62	5.83	4.20	0.38	0.77	3.51

Table 3.6: Mean values for metrics of performance across model formulations.

Table 3.6 on page 71 provides a summary of the metric score for each model specification averaged across all households for each metric of model fit. In other words, each cell in the table is the average of the corresponding metric of performance across 160,000 residences for the corresponding model specification.

Because applications of this work will tend to rely on accuracy of energy prediction, the most important all around metric of fit is the cvRMSE. A comparison of the RMSE and cvRMSE columns illustrates the impact of cross validation. As expected, all cvRMSE values are higher than plain RMSE, but this effect is most pronounced in the second to last model, which has a tendency to over fit the data due to its prescribed change points. It can be seen that the two change-point model on the bottom row edges out the other models for most metrics on average.

Both R^2 and adjusted R^2 reliably reward model complexity. With several hundred observations per model run and only 12 or so model parameters, the model adjusted R^2 values do not differ much from R^2values.

Running counter to other observations, the Kurtosis score tends to be lower for the simplest models (recall that the metric is actually excess kurtosis, where 0 indicates normal error distribution and positive values indicate sharp peaks and fat tails). All model specifications tend to have errors with fat tails, an effect more pronounced as the models improve their other error metrics. This suggests variability that cannot be adequately explained using the regressors on hand, quite possibly the products of unpredictable occupant activity.

By examining the Durbin-Watson and lag1cor metrics, it can also be verified that all models tend to have serially correlated errors, but the more complex models, especially those with a temperature breakpoints, reduce the correlation. It is noteworthy that the metrics of serial correlation for `DOW+toutCP+DL` and `DOW+toutCP+DL+L1` are nearly identical, despite the addition of the lagged temperature term. The correlations overall suggest that households fall into modes of operation that span several sequential days and

the imperfect explanation by the lagged temperature term suggests that very little of the correlation is due to heat storage in the thermal mass of the residence.

3.5.2.1 Best model fits by residence

For each household, we calculated which model produced the best overall value of each metric. For example, the lowest RMSE or the lowest lag1cor or the kurtosis closest to 0. 3.5 shows the count of these "best fits" assigned to each model according to each metric of performance. This visual tool illustrates that the change-point models perform well across most metrics. R^2 and RMSE in particular are dominated by the two change-point model. However, penalizing over-fitting sample data significantly reduces the number of homes best fit by that model. This is consistent with the hint of over-fitting from the cross validated metric averages above. It can also be verified that the lag1cor measure of serial correlation found in the residuals is reduced by structural improvements in the models and the kurtosis of model errors stand out as producing results that seem uncorrelated with the other metrics.

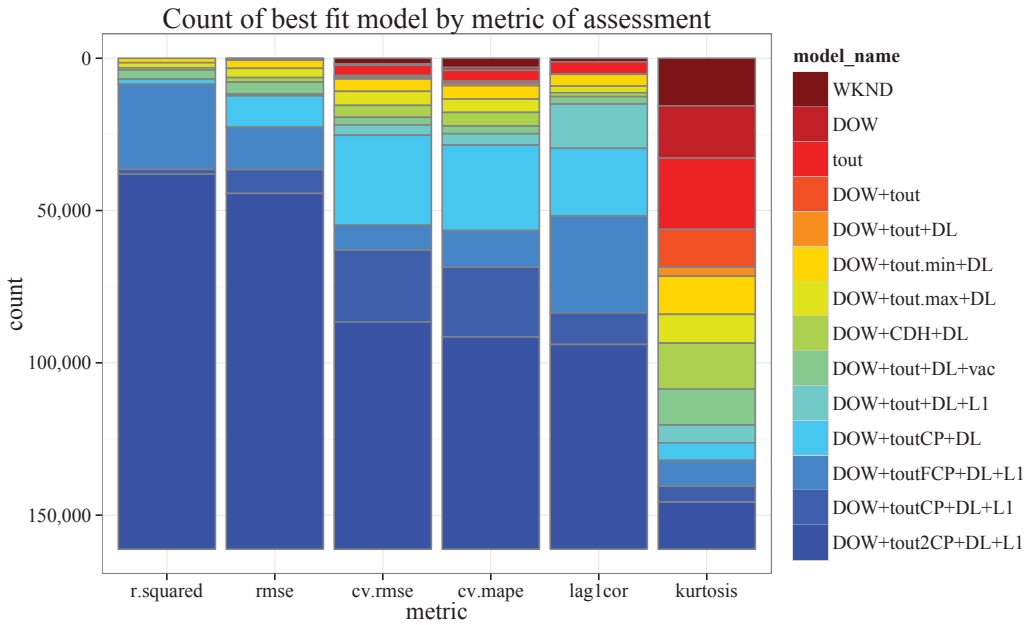


Figure 3.5: Counts of best-fitting models according to different metrics of model performance.

As the view of the counts of best-fit models suggests, subsets of related performance metrics tend to produce similar results. Figure 3.6 on page 73 provides a comparison of the consistency of best fit selections across metrics. Different components of the models improve different metrics, so there is not universal correlation across the metrics.

Four sub-groups with internally consistent outcomes emerge. RMSE, R^2, and adjusted R^2 are metrics that tend to reward model complexity and in-sample fits. cvRMSE and cvMAPE are metrics that measure the quality of out of sample model prediction and penalize over-fit data. lag1cor and DW both address serial correlation, and Kurtosis alone addresses the normality of model errors. It is clear from these correlations that these different dimensions of model performance are somewhat independent.

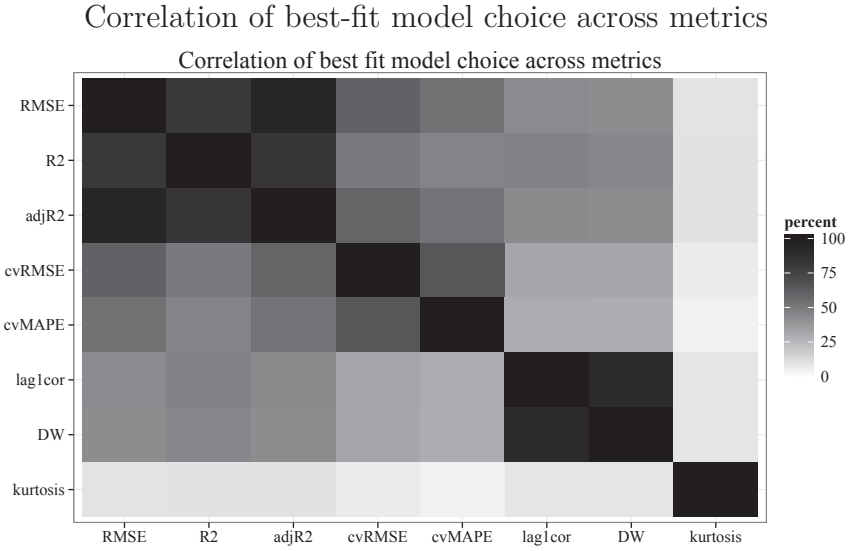


Figure 3.6: Cross correlation of best-fitting by model metric. Diagonals are naturally 100%. Off diagonal terms are darker for metrics that tend to pick the same winners.

3.5.3 Head-to-head model comparisons

In this section, we examine the relative performance of models in head-to-head comparisons. Because we selected regressors with specific patterns of energy consumption in mind, comparisons of models with and without individual regressors are particularly useful. They reveal which households match the patterns described by the regressors and which do not. We begin with visual comparisons and progress to more formal methods of model selection.

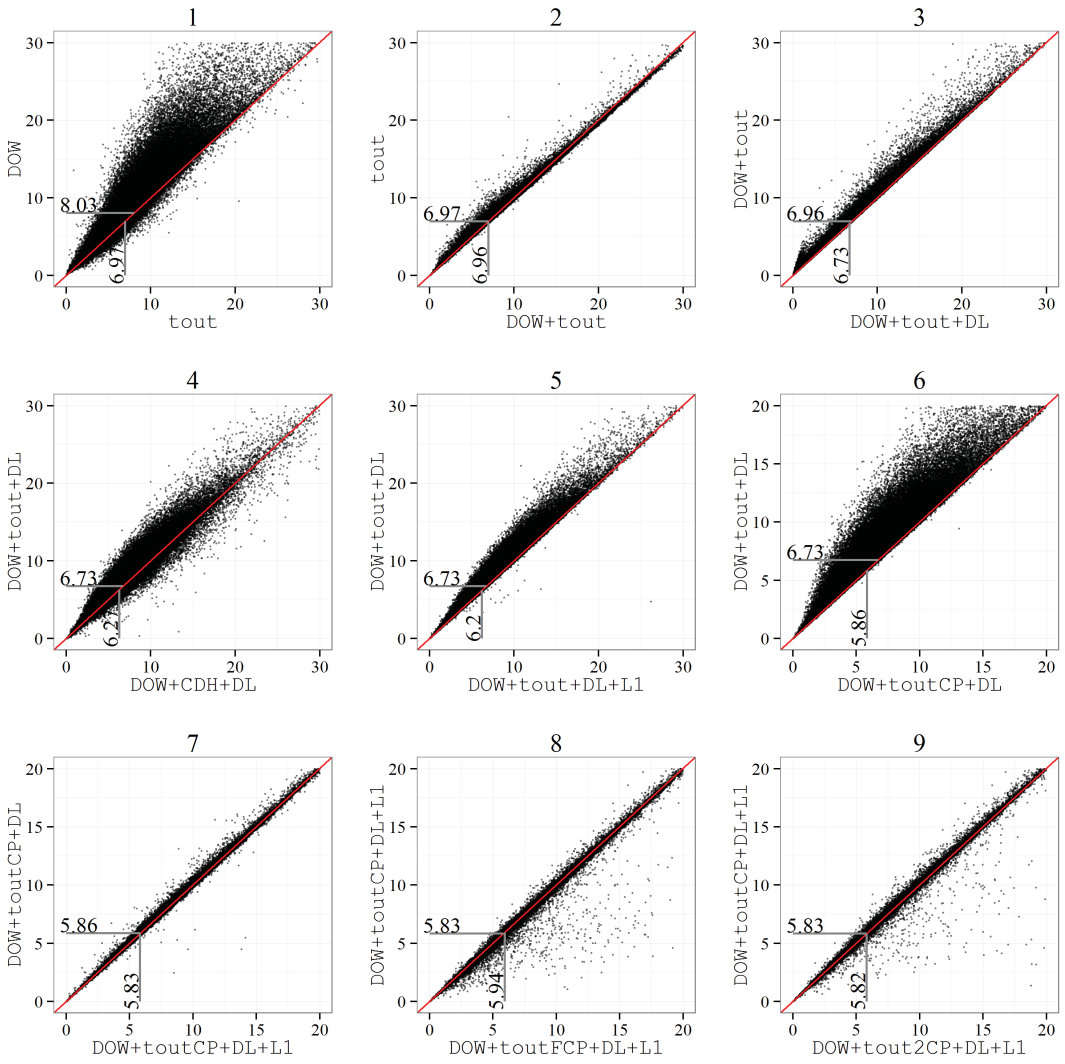


Figure 3.7: Progression of pairwise comparison of model specifications as scatter plots of each household's cross validated RMSE values (in kWh/day) for two models each. The x-axes generally introduce progressively more complex models while the y-axes adopt the previous plot's x-axis. Points on the diagonal lines indicate no change. Points above the diagonals indicate improvement from the new (x-axis) model. Points below the diagonal indicate better fit for the previous (y-model) model. The labeled gray lines are the mean values for each axis.

Figure 3.7 on page 74 uses scatter plots to visualize the change in cross-validated RMSE performance from one model to the next. Each point represents a household's cross-validated RMSE values for the x-axis and y-axis models. Points on the diagonal lines indicate no change. Points above the diagonals indicate improvement from the x-axis

model. Points below the diagonal indicate better fit for the y-axis model. The comparisons are roughly in order from simplest to most complex models. Generally speaking, each sequential plot introduces a new model on the x-axis and the previous plot's x-axis is moved to the y-axis. However, the center row compares the same model to three potential alternatives, and the last two plots compare the same model to two alternatives.

Notice that for most households the model fits improve as the models get more specialized (i.e., each scatter tends to have a majority of points above the diagonal). The `DL` (i.e., day length) term notably improves the model fit (#3), but the most significant gains are associated with the introduction of outside temperature (#1) and a change point in the temperature response, `toutCP` (#6).

Recall that the `toutCP` term is the result of searching for the best temperature value to use for the change point for each household. Notice that it improves upon the plain `tout` term more than the prescriptive `CHD` (#4) and `tout+L1` (#5) terms. Notice also that the single change-point model from #6 has very few points below the diagonal, especially compared to #4.

By fixing the change point and assuming no conditioning below it, `CHD` values can bias model results for residences whose thermal response does not match its assumptions well, causing systematically larger errors with its inclusion (and the large number of points below the diagonal in #4). This is a clear warning that the customary practice of using heating and cooling degree days and hours to control for weather effects in regression models is more risky than often appreciated. They often prescribe their threshold temperature and assume zero thermal response on one side of it. These implicit assumptions can degrade model fits and bias their coefficients.

Building on this concept, we observe that the prescriptive `toutFCP` (#8) term produces a higher average cvRMSE value than the single change point `toutCP` term and has some very notable failures well below the diagonal, despite having more flexibility built into its change points. Where models use their degrees of freedom to over-fit the data in their training sample or where their prescribed change-point locations are extremely inaccurate, their out-of-sample cvRMSE scores can be significantly increased.

Interestingly, the problem of over-fitting affects the overall best performer (as determined by mean cvRMSE) as well. Despite featuring change points that minimize model SSR, the `tout2CP` model (#9) has many points below the diagonal. The three piecewise temperature segments defined by the two change point sometimes over-fit the data by exaggerating their response to outliers. This is especially true for households without systematic thermal responses.

The `L1` term represents a significant improvement when added to a model without a temperature break point (#5), but has a far less significant effect when added to a model with an existing break point (#7). Recall that the `L1` term consists of the previous day's CDD_{65} value and is meant to capture lagged thermal effects. However, CDDs provide a sort of break point at their threshold temperature, and it appears that it is the break point rather than the lag that is responsible for most of the improvement from in #5. The more modest improvement seen in #7 is a more accurate reflection of the value of a

lag in thermal models. This suggests that the dominant thermal lags in most homes are shorter than 24 hours.

Overall, the `toutCP` single change-point model, arguably with the L1 term seen in #7 seems to strike the best balance between good fits and over-fitting.

3.5.4 Formal model selection

The intuition developed through the visual nature of the preceding figures could not have been gained through formal model comparisons. However, formal methods of model comparison can quantify the probability that one model has improved upon another and add rigor to the question of which models fit which households. A traditional comparison of model fit can be done using an F-test, provided the models in question are nested.

For two models m_1 and m_2, with parameter vectors of length k_1 and k_2, respectively, and with k_2 - k_1 = r, the models are nested if there are values for the r parameters that cause the extra terms of m_2 to drop out, leaving behind the exact terms of m_1. For simple linear equations, these parameter values are typically 0. For piecewise regression with different numbers of change points, 2 segments can function as one when their slopes are set equal, creating a single unbroken segment and reducing the number of change points by 1.

However, if the change points don't happen in overlapping positions, the nesting is no longer possible. The single and double change-point models used for this study have completely independent change points, custom fit to each household. Unfortunately, this means that the single change-point model is not nested in the two change-point model. However, if all segment slopes are set equal, we recover an unbroken `tout` term, so `tout` models nest with change-point models.

For $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ the F-metric is:

$$\frac{\frac{(RSS_2 - RSS_1)}{k_2 - k_1}}{\frac{RSS_2}{n - k_2}} \tag{3.4}$$

In other words, it is the ratio of the improvement in RSS per extra term and the RSS of the more complex model (n is the number of observations). If the models are nested, the expectation for the distribution of the F-metric is known, such that p-values for the null hypothesis that the fit of m_2 is not significantly better than m_1 can be computed from the F-distribution, given the degrees of freedom, r and $n - k_2$. Calculated here using R's `pf` method.

Table 3.7 on page 77 is a table of p-values averaged across households from F-tests between more complex model (column) and nested simpler model (row). The numbers in parentheses are the percentage of households with p-values less than 1%, indicating a 99% likelihood that the more complex model has significantly improved the fit. For example, the second row displays the mean p-value and percentage of households for which p < 1% when comparing the `DOW` model to every other model it nests within. True

	DOW	DOW+tout	DOW+tout+DL	DOW+toutCP65+DL	DOW+toutCP+DL	DOW+toutCP+DL+L1	DOW+tout2CP+DL+L1
WKND	0.580	0.069	0.022	0.010	0.008	0.006	0.002
	(11)	(79)	(90)	(95)	(96)	(97)	(98)
DOW		0.057	0.019	0.010	0.008	0.006	0.003
		(81)	(91)	(95)	(96)	(96)	(98)
DOW+tout			0.109	0.025	0.023	0.014	0.005
			(64)	(90)	(90)	(92)	(95)
DOW+tout+DL				0.086	0.078	0.042	0.017
				(74)	(75)	(80)	(88)
DOW+toutCP+DL						0.152	
						(37)	

Table 3.7: Mean p-values for F-test comparisons between more complex model (column) and nested simpler model (row). The p-value should be interpreted as the probability that the more complex model has not actually improved upon the simpler one, so smaller values are indicative of greater improvements. The parenthetical values are the percentage of households with p-values less than 0.01 for the relevant model pair.

to the limitations of the test, only nested models have been compared in this way. The differences in locations of change points between `toutCP65`, `toutCP`, and `tout2CP` prevent nested comparison between them.

The AIC metric, defined in Table 3.5 on page 66, is derived from information theory treatment of the likelihood of observing the data actually observed, given the model predictions. Because it is not applied as a formal test statistic, [17] documents that AIC has the virtue of being able to compare model fits regardless of whether they are nested. Smaller AIC values are better, but the metric contains a term that increases with the number of model parameters. So AIC is able to balance model complexity against goodness of fit. Table 3.8 on page 78 is a compilation of the percentage of households for which the column model has a lower AIC value than the row model.

For both the F-test and AIC metrics of model comparison, the lessons are largely consistent with the qualitative findings from the preceding comparisons of model performance. Our carefully selected model additions are generally worth the added complexity. The change-point models in particular out-perform the others. One surprise was how consistently both comparison metrics rejected the benefits of moving from a weekday/weekend indicator variable to 7 day of week indicators. The addition of 6 new regressors in one step is a large change in model complexity. For most households, the difference in consumption within weekdays must not be significant enough to justify their inclusion on statistical grounds.

The two comparisons disagree on the merits of adding an `L1` term (recall that this

	DOW	DOW+tout	DOW+tout+DL	DOW+toutCP65+DL	DOW+toutCP+DL	DOW+toutCP+DL+L1	DOW+tout2CP+DL+L1
WKND	18.0	85.8	94.4	97.2	97.4	99.0	99.5
DOW		90.3	96.5	98.0	98.2	99.6	99.8
DOW+tout			81.1	95.4	95.5	99.5	99.8
DOW+tout+DL				85.2	86.0	98.9	99.1
DOW+toutCP65+DL					75.7	94.8	99.2
DOW+toutCP+DL						99.0	98.7
DOW+toutCP+DL+L1							79.8

Table 3.8: Percentage of households where the AIC metric chooses each column model as better that the row model. Because AIC does not require models to be nested, it allows for comparisons across all models.

CDD_{65} from the prior day) to the `DOW+toutCP+DL` model. We have observed that the `L1` term tends to interact with the upper temperature segment slope coefficient estimates, increasing standard errors and increasing the probability of the fit producing a zero or negative slope. In other words, due to consistency of weather from one day to the next, the `L1` term and upper temperature slope are often co-linear, so adding the `L1` term is not always worth the added complexity.

As with other metrics of fit, the two change-point `DOW+tout2CP+DL+L1` model tends to very narrowly outperform the other change-point models in the F-test and AIC tables. However, its AIC comparison with the single change-point `DOW+toutCP+DL+L1` model (in the the lower right of Table 3.8 on page 78) produces one of the weakest results in the table. This is also consistent with the weak improvement and occasional degradation in performance seen in plot 9 of Figure 3.7 on page 74 and with the interpretation that the two change-point model will often over fit data for modest gains in performance.

3.6 Results: Physical interpretation and applications

Because the models were designed with physical interpretation in mind, their coefficients are of significant interest. This section examines what can be learned about the building stock and individual homes from the regression coefficients.

3.6.1 Inferred residential characteristics

Based on the evidence of overfitting in the two change-point model and the ambivalence of the model comparison test statistics on the value of the `L1` term, we have selected the parsimonious `DOW+toutCP+DL` model for applied analysis in this section. As we have seen, the additional degrees of freedom in more complex models can lead to spurious fits and colinearity that confound physical interpretation of model coefficient values.

Figure 3.8 on page 80 summarizes the coefficients for the major regressors of the `DOW+toutCP+DL` model. The day length term in the model has been interpreted in terms of lighting energy and the individual DOW regressors have been averaged to produce a single average estimate of daily baseline energy, which is the expected daily energy usage absent heating, cooling, and day length effects. The lower and upper temperature slopes refer to the regression fits for the heating (lower) and cooling (upper) thermal response across the change point.

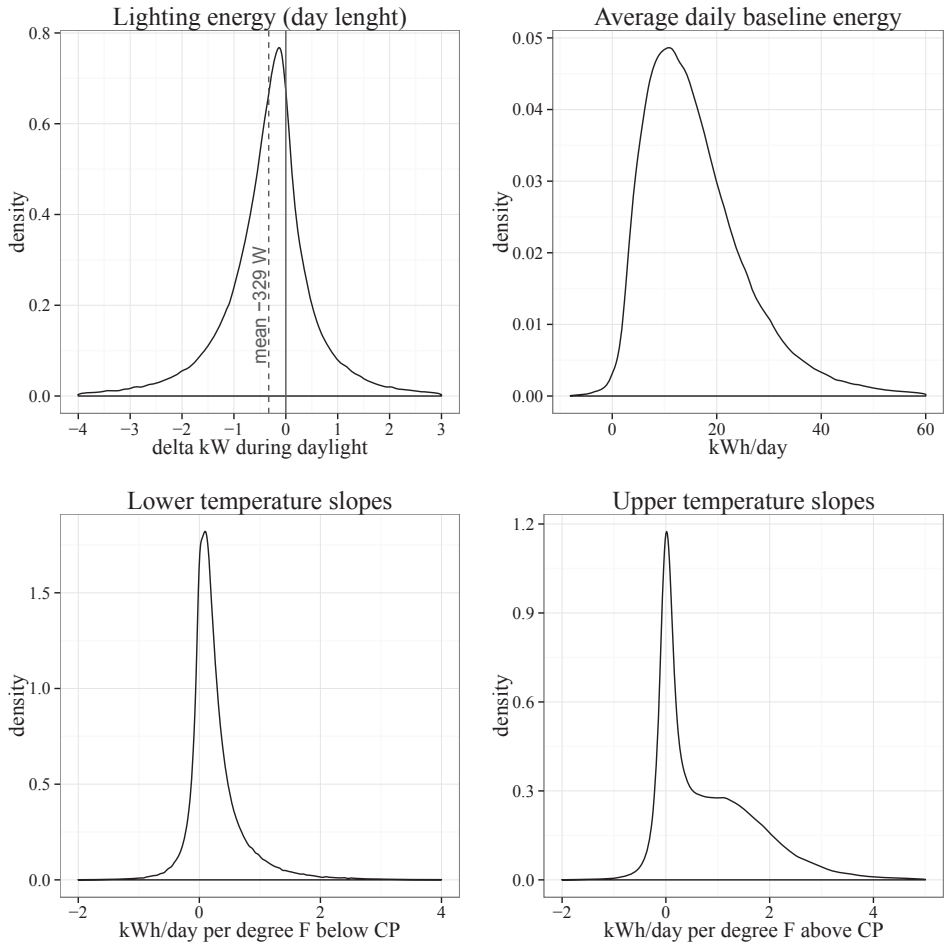


Figure 3.8: Summary of the coefficients for `DOW+toutCP+DL` model fit across 160,000 households.

The distribution in the upper left is based on the distribution of day-length regression coefficients. Recall that the day-length regressor is the number of hours of daylight for each day minus the hours of daylight on the winter solstice. This means that the day-length coefficient has units of kWh/hr of daylight each day and corresponds to the expected change in daily energy use for every additional hour of daylight. Both the mean and peak of the distribution are negative, implying systematically less energy use during daylight hours.

The most logical explanation for this distribution is that some lighting energy is avoided when days are longer. We might worry that day length is correlated with season and that seasonal weather drives conditioning loads. However, longer summer days are associated with air conditioning, so we would expect the effect to have a positive, not negative sign. Although there is significant variability, most likely caused by other systematic changes in non-thermal seasonal energy consumption, the distribution indicates that extra

hours of darkness produce additional demand averaging about 330W per household.

The daily baseline energy distribution in the upper right is the expected daily energy use after controlling for thermal and day-length effects. It can also be thought of as an estimate of the energy used by all appliances and electronics in regular use in households. Despite its it modest sounding definition, it is typically responsible for most of the energy used by a household. Data from [31] suggest that 66% of the annual energy use of the average California residence falls into this category.

The bottom two distributions provide coefficient values for the lower (heating) and upper (cooling) modeled temperature response segments. The distribution of heating slopes is fairly uniform, with a positive mean that is just above zero and a positive tail somewhat fatter than the negative tail. This is unsurprising because most heating in PG&E's territory makes use of on-site combustion. The electric portion of these heating systems is restricted to fans, pumps, and other auxiliary loads.

The shape of the cooling-slope distribution is consistent with overlapping distributions for homes with and without air conditioning. The estimates for homes without air conditioning should have a mean of zero, with some error above and below true zero. The large peak at zero in the cooling slope distribution fits this description very well.

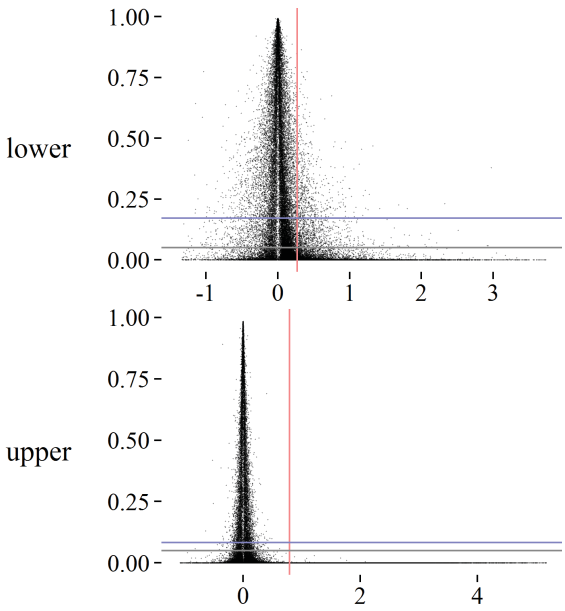
The estimates from homes with air conditioning should have a positive mean and a fairly broad distribution because the coefficient values will be influenced by different home sizes, levels of insulation, equipment performance, and occupant preferences. The broad plateau and gradual decline in the densities of slopes around 1kWh/day per degree above the change point and higher fits this description well.

3.6.1.1 Fraction of households with electric heating and cooling

We know that the heating and cooling coefficient distributions represent households with and without conditioning, so we interpret them as two overlapping independent distributions. Here we estimate the percentage of homes with different configurations of heating and cooling. This can be done using model coefficients and p-values. Figure 3.9 on page 82 shows scatter plots of lower and upper temperature segment coefficient values on the x-axis vs. the corresponding p-values on the y-axis. The red line is at the coefficient average, the blue line is at the p-value average, and the gray line is at the p-value 0.05. We take positive model coefficients with $p<0.05$ or $p<0.01$, as indicators of heating or cooling loads and assume that all other values are indicators of no heating or cooling loads. We can then use these indicators to compute the percentages of homes in each class of heating and cooling. These percentages are displayed for both p-value thresholds in the table to the left of the plots.

These estimates suggest that 68-70% of households have some form of cooling and 50-57% have some form of electric heating. For reference, RECS 2009 data suggests that 57% of California homes have electric cooling and 38% have electric heat. In both cases, we have over estimated. It is likely that the lowest coefficient values are not true heating and cooling responses. Refrigerators work harder in the summer and many gas furnaces

have electric blowers. These can create non-zero temperature responses that meet our criteria for selection.



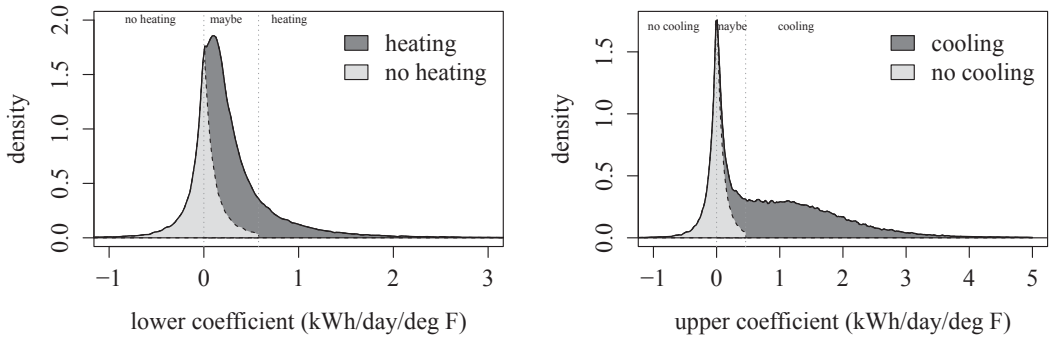
Fraction of residences in each cooling class

	both	cooling only	heating only	neither
p<0.05	42	28	15	15
p<0.01	36	32	14	18

Figure 3.9: Coefficient value (x-axis) vs. p-value (y-axis) scatter plots for the lower and upper temperature segments of the `DOW+toutCP+DL` model. The red line is the mean of the coefficient values. The blue line is the mean of the p-values. The gray line below it is set to a p-value of 0.05.

Another approach to estimating heating and cooling percentages comes directly from the distributions of coefficient values in Figure 3.10 on page 83. We know that the heating and cooling distributions should be mixtures of the mean-zero symmetric distributions of homes without electric conditioning and positive mean asymmetric distributions of homes with electric conditioning. We also know that the coefficients associated with heating and cooling should not take on negative values. Combining this information, we conclude that the negative portion of each distribution of coefficient values is the negative side of the symmetric mean zero non-conditioning distribution. For both heating and cooling, we mirror the negative portion across the zero line to estimate the full, symmetric, no-conditioning distributions. The percentages of residences within these distributions is an estimate of the percentage of homes with no electric conditioning. The remaining fraction is the percentage with electric conditioning.

Illustrations of the estimated fraction of households using electric heating and cooling



	no cool	maybe cool	yes cool	**total**
no heat	6.3%	5.1%	9.0%	**20.4%**
maybe heat	9.4%	19.2%	36.4%	**65.0%**
yes heat	4.1%	4.9%	5.5%	**14.5%**
total	**19.8%**	**29.2%**	**50.9%**	

Figure 3.10: Illustrations estimating the number of households with electric heating and cooling within the total distributions of temperature response coefficients and a table of definite no, maybe, and definite yes percentages of households for both heating and cooling.

This process is equivalent to doubling the empirical CDF value of each distribution at 0 to find the non-conditioning percentages and then subtracting that value from 1 to find the conditioning percentages. Using this method we estimate 60% of households have electric cooling and 59% electric heating. Figure 3.10 on page 83 illustrates the method. We can also define a threshold past the 99th percentile of the no-conditioning distributions as almost definitely involving conditioning. This segments the distribution into three zones. Below zero, we assume there is *no* conditioning. Between zero and the 99th percentile threshold, we assume there *may be* conditioning and there may not be, and above the threshold, we assume that *yes* there is conditioning. The percentages of households in these zones are also provided in the figure. The numbers associated with yes/maybe categories capture the underlying uncertainty in the heating estimates and the robustness of the cooling estimates.

3.6.1.2 Annual heating and cooling energy

Because the `DOW+toutCP+DL` model is structured to estimate temperature response terms relative to the change point, total annual conditioning energy can be estimated summing the multiple of the coefficients and daily average temperatures for a full year for each household. These calculations are expressed in equations (3.5) and (3.6).

$$kWh_{cooling} = \sum_{d \in year} \beta_{tout+}(tout_d - CP)_+ \tag{3.5}$$

$$kWh_{heating} = \sum_{d \in year} \beta_{tout-}(CP - tout_d)_+ \tag{3.6}$$

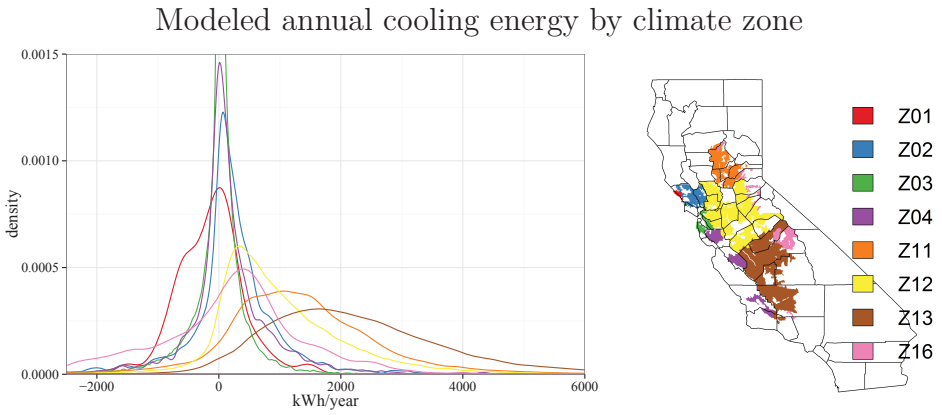


Figure 3.11: Annual cooling energy

Figure 3.11 on page 84 presents the results of the household total annual cooling energy calculations in one distribution per climate zone. Zones 11, 12, and 13 in are the hottest climate zones in the territory and very clearly tend toward higher estimated annual cooling.

The annual cooling energy data can also be used to create maps of average residential heating and cooling energy by zip code, as seen in Figure 3.12 on page 85. The pattern of cooling energy is quite clear, with modest cooling demand on the coast and in the mountains and higher demand in the Central Valley. The pattern of heating energy is less obvious. This is partially because most heating energy comes from on-site natural gas combustion, not electricity. Perhaps the hot spots on the map are locations with unusual numbers of electric heaters, or with higher than usual penetration of heat pumps.

Maps of mean modeled annual heating (left) and cooling (right) energy by zip code

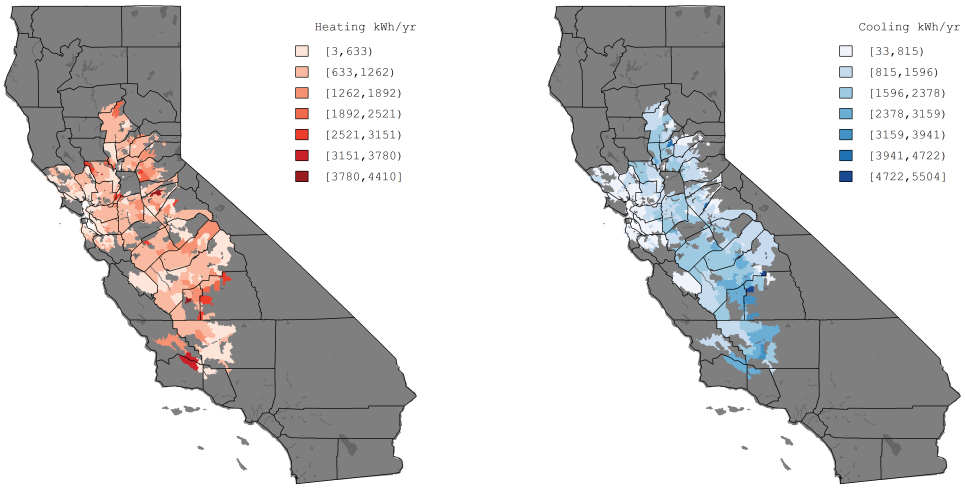


Figure 3.12: Map of the electric portion of annual heating (left) and cooling energy (right) according to the `DOW+toutCP+DL` model.

Totaling the residential cooling energy of our sample and scaling it to the 4.5M customers of the service territory produces rough estimates of total annual conditioning demands in the PG&E territory[5]. However, it is clear that a significant portion of the heating and cooling coefficients have negative values. These correspond to apparent heating at higher temperatures and apparent cooling at lower temperatures. If we assume that these are spurious results and that the distribution of spurious results is symmetrical across 0, the negative estimates will cancel out the spurious positive results as they are summed. The naive practice of summing across all results becomes the equivalent of removing the contribution from model outcomes presumed to be spurious. This can be considered a conservative estimate of conditioning demand. If all negative results are simply ignored by the sum, a larger, less conservative, estimate is produced.

Together, these methods provide estimates of 4370-4790 GWh cooling energy per year and 1940-2110 Gwh heating energy per year. These correspond to 14.0-15.4% (cooling) and 6.2-6.8% (heating) of annual total residential electricity sales[6]. However, it should be cautioned that end uses like water heating and pool heating will be lumped in with space heating in these results. Similarly, refrigerators, which work harder in warmer weather, will be lumped in with the cooling results.

It is instructive to compare these results to estimates derived using other methodologies. RECS 2009 from [31] Table CE4.10, allocates 18% of total California annual kWh to cooling and 15% to heating. California's Residential Appliance Saturation Survey

[5]We caution that the stratified sample was not fully random, so the sample is not technically fully representative.

[6]PG&E sold 31,200 GWh to residential customers in 2009.

(RASS) from 2009, as described in [51], allocates 7% of total California annual kWh to cooling and just 2% to heating. Our ranges of estimates for just the PG&E territory fall between RASS and RECS values, which is an affirmation that the models are recovering information within the range of expected aggregate values. However, the disagreement among all methods suggests that more work is needed to explain the discrepancy among the bottom-up (RASS), top-down (RECS), and empirical regression approaches.

3.6.1.3 Day-of-week effects

The day-of-week coefficients in the model specification effectively allow separate estimates of expected levels of energy consumption after controlling for thermal loads and day length. At the level of individual homes, these estimates will be indicative of prevailing patterns of weekly occupancy and energy use. Averaged across all homes, these values provide an indication of prevailing patterns of energy use.

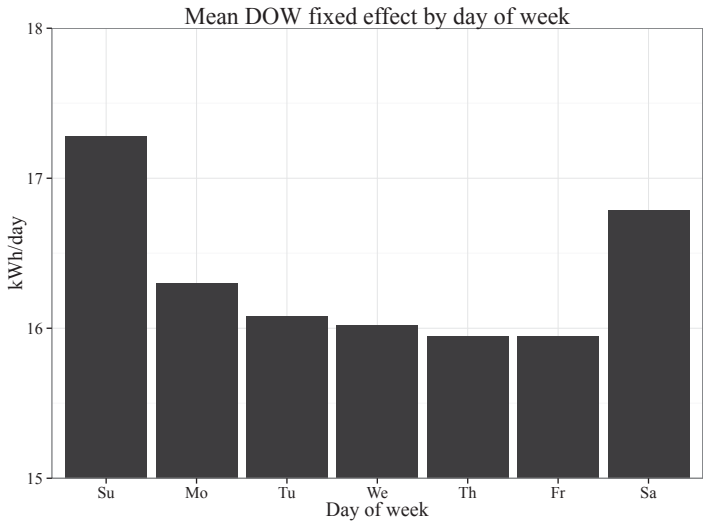


Figure 3.13: Day of week coefficients from the `DOW+toutCP+DL` model, averaged across all homes.

Figure 4.13 on page 120 presents the average for each of the 7 day-of-week fixed effects across all households. Note the scale of the y-axis, which does not reach 0. Across all households, Saturday and Sunday involve greater expected energy use than weekdays. The difference is on the order of 1 kWh/day, easily achieved through greater appliance and plug load use, such as a few hours of watching TV, doing a load of laundry, or hosting friends for dinner. Among the weekdays, expected energy use is higher early in the week and lower toward the end of the week. This may reflect a population-wide preference for more out-of-the-home social activity as the week progresses, or it may be an artifact of preferences for major appliance usage for things like laundry and dish washing.

3.6.1.4 Change-point estimates

Although change points are estimated via their own separate model fitting process, they are still model parameters that contain information about household energy use. Figure 3.14 on page 87 visualizes the change points found for the `DOW+toutCP+DL` model as a function of mean summer outside temperatures. The means for each temperature band are stable around 65°F, with the inter-quartile range less than 10°F. Because there are other heat gains in homes (including solar gains and waste heat from energy consumption and people), this is consistent with expectations about what outside temperatures will lead to indoor cooling. The cooler the change point is, the greater the possibility there is energy being wasted on an indoor setpoint that is cooler than necessary or on neutralizing avoidable thermal gains.

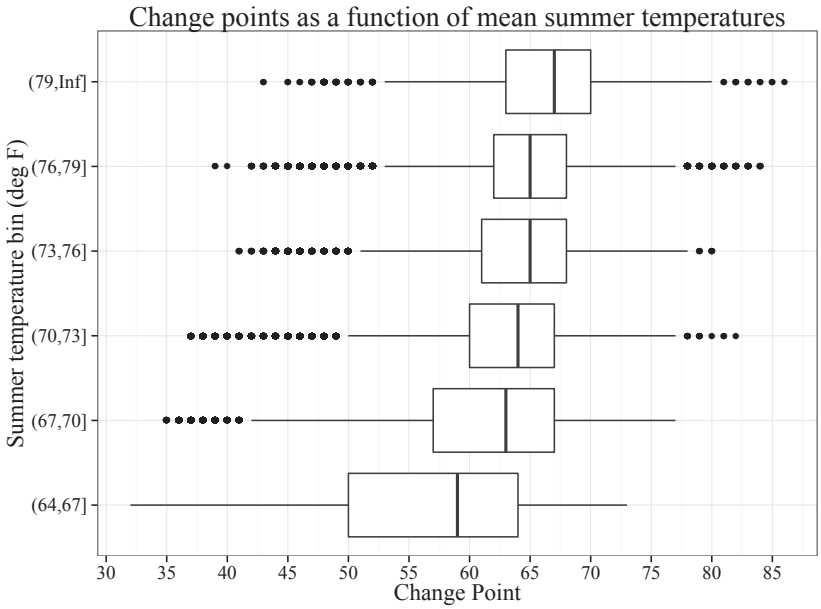


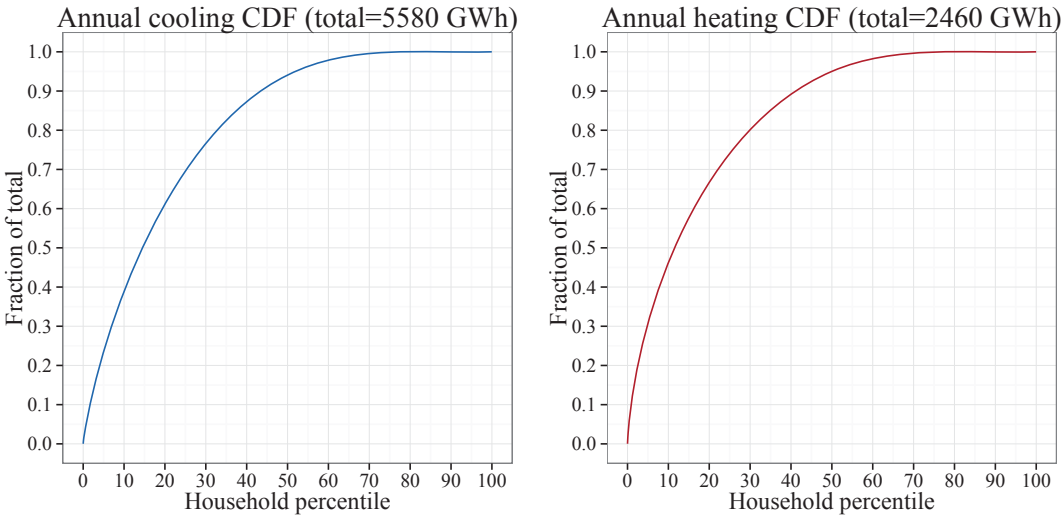
Figure 3.14: Change-point distributions for the CPs from the `DOW+toutCP+DL` model within bins of mean summer (May through September) daily temperatures. The center line of each box plot is the mean, the rectangles span from the 25th to 75th percentile, the whiskers extend by 1.5 times the inter-quartile range in either direction, and the points are all the observations beyond the whiskers.

Change points appear to drift slightly upward with hotter outdoor temperatures. This is consistent with better insulation and tighter construction in areas that really feel the heat. It is also consistent with people adapting their physiology, expectations, or clothing to climate conditions. As [24] show, the temperatures deemed neutral by building occupants will increase as they are consistently exposed to warmer conditions in naturally ventilated buildings.

The modest increase is also consistent with a potential biasing in the algorithm that chooses the change points. The hotter the climate the more points there will be above the change point and the stronger the incentive to fit those points well. The lowest band of summer temperatures underscores the fact that the change-point algorithm always returns a change point that improves model fits, but the cooler it gets the less likely those points are to be connected to actual cooling set points.

3.6.2 Differentiation of homes

With an eye toward improving efficiency and demand response program outcomes, it can be of great practical value to uncover the signatures of high levels of consumption, waste, and inefficiency in specific homes, and to provide them with targeted messages and programs designed to correct their problems. This section provides three examples of how that can be done with the modeling approach taken by this study. The first differentiates homes using estimated annual heating and cooling energy. The second differentiates homes based on estimated changes in annual energy caused by set point changes. The third differentiates homes based on their peak system demand day coincident consumption.



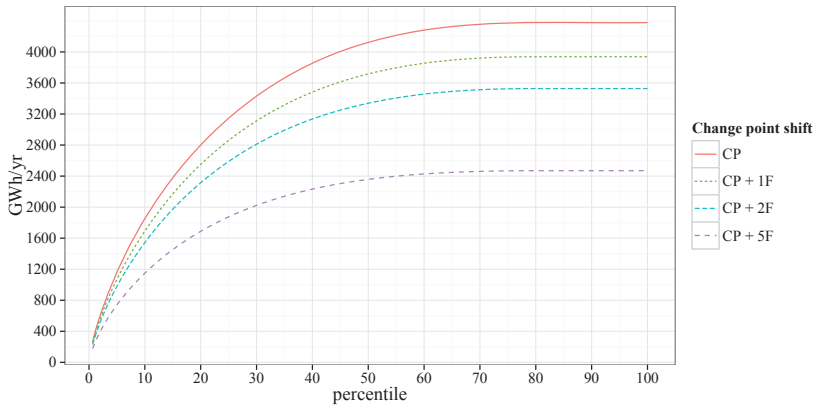
percentile	5th		10th		25th		50th		75th		100th	
	GWh	%	GWh	%	GWh	%	GWh	%	GWh	%	GWh	%
cooling	1315	24%	2166	39%	3881	70%	5247	94%	5574	100%	5579	100%
heating	755	31%	1136	46%	1820	74%	2336	95%	2457	100%	2459	100%

Figure 3.15: Cumulative distributions of annual electric cooling and heating demand within the sample, with a table of corresponding cumulative totals by percentile of household.

Figure 3.15 on page 88 presents cumulative distributions of annual cooling and heating energy, as calculated using equations (3.5) and (3.6), across the households in the sample. To provide ballpark estimates of total potential, the stratified sample totals have been

scaled in the figure titles to values that represent PG&E as a whole. These cumulative distributions support calculations of the relative benefits of targeting HVAC efficiency programs to the right households. As can be verified, 50% of annual cooling, corresponding to over 2000 GWh in the pattern is typical for the whole territory, comes from just 15% of residential customers, while 90% of cooling comes from 45% of residential customers. The final 10% of cooling energy comes from 25% of residential customers, and 30% don't appear to cool at all. A program that achieves a fixed-percentage improvement in cooling efficiency for a fifth of the sample population will return 3x the benefits if it is targeting to the top consumers instead of selecting randomly from within the population. Similar reasoning applies to heating energy.

Cumulative distributions of annual cooling energy with modified cooling set points



Savings from set point increases by ranked percentile of the population

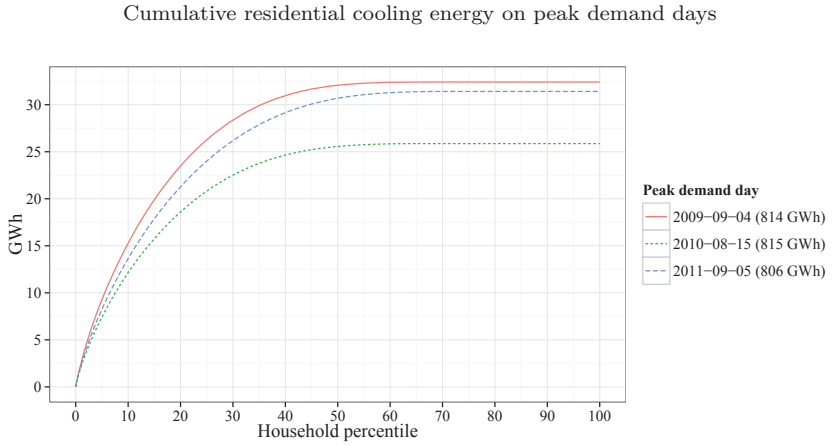
percentile	5th		10th		25th		50th		75th		100th	
$\Delta°F$	GWh	%	GWh	%	GWh	%	GWh	%	GWh	%	GWh	%
$+1°F$	95	8%	157	8%	289	9%	406	10%	440	10%	440	10%
$+2°F$	185	16%	306	17%	561	18%	784	19%	849	19%	850	19%
$+5°F$	425	36%	702	38%	1276	40%	1766	43%	1906	44%	1909	44%

Figure 3.16: Cumulative distribution (top) of annual cooling energy extrapolated from the stratified sample to the full service territory, subject to an increase in thermostat set points by 1, 2, and 5°F, with a table (bottom) of corresponding annual cumulative cooling energy savings values by percentile of the population.

The regression model fits, complete with change points and temperature sensitivities, offer an excellent opportunity to calculate the savings potential from a common efficiency prescription: increasing a household's cooling set point. By altering the change point used by each residence and re-calculating the `DOW+toutCP+DL` model predictions of cooling energy consumption, we can estimate an altered annual total of cooling energy for every residence. Figure 3.16 on page 89 presents the cumulative annual energy associated with residential cooling and extrapolated from the sample to the full service territory, sorted from highest consuming residence to lowest, for the unaltered change point, and the change point increased by 1, 2, and 5°F. This exercise reveals that a universal set point

increase of 5°F would reduce cooling energy demand in PG&E's territory by over 1900 GWh/yr, assuming the sample is representative of the territory. This is greater than a 40% decrease in annual cooling energy and equal to approximately 6% of all residential sales.

This finding is consistent with [44], which altered cooling setpoints in simulation models and found that a 1°C (1.8°F) increase saved 7-15% of cooling energy and a 4°C (7.2°F) increase saved 35-45%. Set point changes offer a no cost and effective way to reduce cooling loads, but implementing such changes has traditionally relied on significant behavior modification by residents. The latest generation of smart and communicating thermostats offer a renewed opportunity to reduce cooling energy with more passive participation by customers.



Cumulative peak demand day cooling energy by ranked percentile of the population

percentile	5th	10th	25th	50th	75th	100th
	GWh	GWh	GWh	GWh	GWh	GWh
2009-09-04 (814 GWh)	9	15	26	32	32	32
2010-08-15 (815 GWh)	7	12	21	26	26	26
2011-09-05 (806 GWh)	8	14	24	31	31	31

Figure 3.17: Residential cooling contribution to peak demand days, with table of cumulative residential cooling by household percentile.

Targeting can also be applied to understanding how to best shave peak demand. To do this, we identified the days with highest demand using historical load data maintained by CAISO. We will call this set of days *dpeak* and a particular one $dpeak_i$ We then matched temperature data for every zip code, *zip*, to the dates of interest: $tout_{zip,dpeak_i}$. Then, using the change points and temperature slopes associated with every household in each zip code, we calculated the expected daily energy demand per household, h:

$$kWh_{h,dpeak_i} = \beta_{tout+_h}(tout_{zip,dpeak_i} - CP_h)_+$$

By sorting the values of the sum from the largest to smallest and calculating a cumulative sum, we produced cumulative distributions of residential cooling energy on peak demand

days, extrapolated from the sample to the full service territory. Figure 3.17 on page 90 provides an illustration of these cumulative curves, scaled from our sample size to the entire PG&E territory, for the peak days of 2009, 2010, and 2011. Using these curves, we can produce a list of the best individuals to target for maximizing energy savings on peak demand days. For example, 50% of the cooling energy consumed by residences in the PG&E territory during those days can be attributed to 11-18% of households, depending on the day.

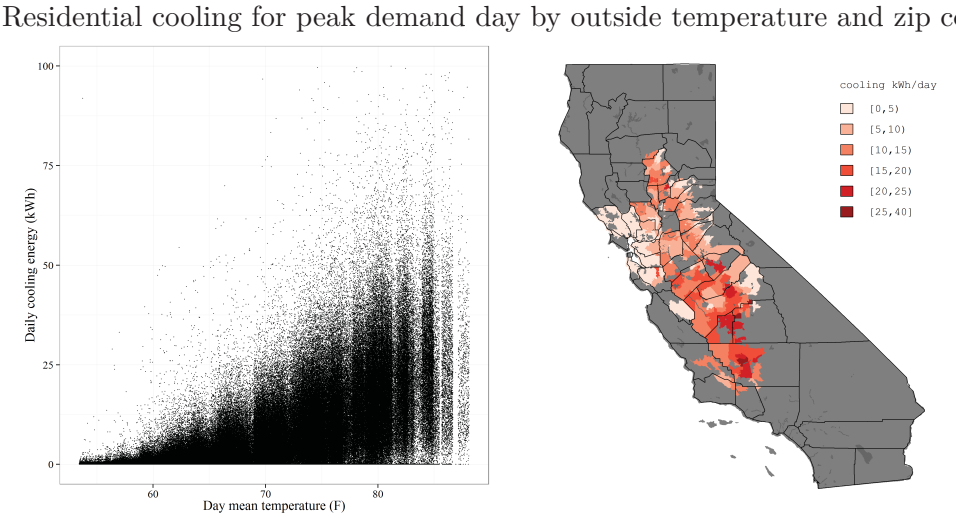


Figure 3.18: Residential cooling for peak demand day scattered against outside temperature (left) and mapped.

Because the peak day load calculations are done separately for each household, we have a very rich set of information on where consumption happens on peak days. Figure 3.18 on page 91 visualizes the daily cooling demand on peak days by outside temperature on the left and maps the same data on the right. The scatter plot illustrates how strong the influence of temperature is over peak day demand. However, it also shows very significant variation within temperature bands. This variation is surely due to many different factors, with home size, solar orientation, air tightness, insulation levels, cooling equipment performance, set points, and occupant preferences and behaviors prominent among them. A better understanding of these factors could be used to better target peak demand savings programs and inform residents about where they stand relative to their peers.

The map of cooling demand echoes the correlation between hot weather and high consumption. However, the location of the demand is instructive. The Central Valley produces much higher mean household demand than the coast or hills. Programs targeted at the zip code level within this area could readily deliver greater savings for less money than similar programs in other areas.

3.7 Conclusions

The conclusions drawn from this work are separated into three categories. First, we comment of what has been learned about the energy use categories, introduced in Table 3.2 on page 56, that were used to motivate and structure the semi-physical regressors and models used in this study. Second, we draw conclusions about the models and modeling process. Finally, we summarize what can be learned by interpreting the modeling results in the context of utility program planning and implementation.

3.7.1 Energy use categories

Table 3.9: Summary of results of inquiry into household characteristics identified at the beginning of this work.

Name	Description
Thermal response	Using the most generous range possible, we estimate 51%-80% of trhe sampled households have some form of air conditioning and 15%-80% have some form of electric heating. Most of these loads come from HVAC systems, but refrigerators, electric hot water heaters, hot tubs, and heated pools are also temperature responsive.
Change points	Change points in thermal response are critical to improving the performance of regression models of home energy use. However, a prescriptive point that is too high, too low, or inappropriate can introduce bias into the model fit and will impact other coefficients, especially the temperature response terms and the estimates of day of week energy. Custom estimates of change points fit households the best. Derived change points average around 65°F (18°C) but tend to drift upward at higher temperatures. The spread in estimated change points suggests major opportunities to save energy with zero cost adjustments to thermostat set points. A universal increase of 2°F for all homes is extrapolated to the state level is projected to save the PG&E service territory 850 GWh/yr.

Thermal lags	Daily data benefits only slightly from the inclusion of a lagged temperature term. This suggests that the strategy of working with daily data has effectively addressed concerns about thermal lags confounding model results. However, this also means that this approach cannot do much to estimate the strength and time scales of such lags. Hourly data can be expected to include lagged responses, but these effects will need to be disentangled from occupancy, scheduling, and solar geometry, which can coincidentally produce effects that lead or lag air temperature. More work is needed on lags in data at finer timescales.
Lighting	The consistency of response to the day-length regression term and the sign of the mean and majority of day-length coefficient fits strongly suggests that a lighting effect is present in the data. Longer days are in the summer, and should therefore have higher cooling loads and higher daily consumption. The fact that the prevailing impact of day length is negative (longer days reduce energy use when other effects are controlled for) suggest that home are using less lighting energy during longer days. The mean of this effect suggests average lighting load of about 330W. However, other seasonal changes in patterns of energy use could also explain this feature. Further work with hourly data and ground truth data is needed to determine how accurately the day length term captures lighting demand.
Minimum loads	Typical daily energy, but not minimum intensity, can be derived from regression outcomes of daily data. The typical day of week values often dominate home energy use, so the contributors to this consumption deserve further scrutiny. Assuming that there are days without any consumption other than the minimum loads, the lowest daily total could be taken as proxies for 24hrs x minimum power. However, true minimums are best addressed with hourly data.

Occupancy	Hourly occupancy is not detectable using daily models, but a day of week effect is present in the data. A majority of homes show increased energy usage over weekends, which is consistent with higher occupancy on days where occupants are off work. Homes where this effect dominates can be distinguished from homes where it does not — some people may work non-traditional weekly schedules, have one or more household member home most days, or spend significant time away from home on weekends. Occupants are likely suspects for producing the fat tails in model residuals. The unpredictability of human behavior is also be present in the households whose fits are poor, no matter the structure of the model attempted. Finally, we might think of occupant loads, scheduled loads, and thermal loads as independent processes all talking place in a home. If this is the case, homes without cooling should provide insights into what is happening with occupants in homes where cooling masks parts of their behavior.

3.7.2 Models and modeling process

There are many reasons and ways to model residential energy use, and the academic literature provides examples of a great variety of approaches from different disciplines. Because our work is exploratory in nature and based on a large data set, we anticipate that our findings will be relevant to many of these. However, the lessons learned during the course of our modeling efforts are most applicable to the common practice of weather-correcting energy demand data, the evaluation of energy efficiency program impacts, and the estimation of heating and cooling demand.

In general, we found that there is no single model that provides the best fit for all households. The variety of best-fit models provides a window into the diversity of patterns of consumption found in households. However, models that include a piecewise response to outside temperature and seasonal varying day-length terms perform best on average. Even with penalties for model complexity and over-fitting data, these model features tend to improve model fit metrics.

Of all the regressors examined, change-point temperature response terms provided the greatest single improvement in model fits. A term that captures variation in day length reliably improves model fit and is consistent with interpretation in terms of lighting energy. However, this term is capable of explaining other seasonally variable patterns of consumption, and there is substantial variability in coefficient fits. A thermal lag term can improve model fits but tends to interact with unlagged thermal response coefficient fits and may not be worth including. Typical residential buildings are not massive enough to have thermal lags of over 24 hours.

We have demonstrated that the customary practice of using heating and cooling degree days or hours to control for weather effects in regression models is more risky than often

appreciated. This is because they prescribe their threshold temperature and assume zero thermal response above (heating) or below (cooling) it. These implicit assumptions can degrade model fits and bias their coefficients.

Homes with large cooling loads tend to be more predictable (as a percentage of variance) than those with smaller cooling loads. This suggests a model of consumption with predictable engineered equipment overlaid on top of the underlying less predictable behaviors of people.

The distribution of model fits in the population reveals some households that are poorly fit by all of our model specifications. These households follow their own unobserved internal logic. The models define patterns of consumption that these households *do not follow*, which can still be useful information to have in the context of program targeting.

Finally, we observe that different metrics of model fit are responsive to different modeling concerns. Fit metrics should be carefully selected for applicability to the question being studied. For many applications, the percentage of variance explained or metrics of predictive accuracy can be inappropriate or incomplete indicators of applicability. In our case, we were most interested in the profile of changes across metrics that followed the addition of specific regressors to models.

3.7.3 Applications

The most practically useful outputs of this work are the estimates of semi-physical characteristics of 160,000 households drawn from a stratified sample of PG&E's service territory. We have estimates for the expected daily energy use by day of week and in response to day length and outside temperature. Our temperature fits include estimates of heating and cooling responses, with associated change points. These terms have predictive power and logical physical interpretations, so we have used them to estimate several useful, but previously unobserved, characteristics of residences in PG&E's territory.

Based on coefficient values and their corresponding p-values, we estimate 68-70% of households have some form of air conditioning and 40-57% of households have some form of electric heating. RECS 2009 survey data suggests that 57% of California homes have electric cooling and 38% have electric heat. Our estimates likely include the temperature sensitivities of refrigeration and electric water heating and the electric energy associated with furnace fans. In other words, out method of estimation is sensitive to small correlations between energy and outside temperature not primarily driven by heating and cooling.

By calculating annual heating and cooling consumption for each household and summing across households, we obtain totals corresponding to 14.0-15.4% (cooling) and 6.2-6.8% (heating) of annual total residential electricity sales. The top 10% of residences were responsible for 39% of cumulative annual cooling energy and the top 25% were responsible for 70%. The lowest 50% of households contribute just 6% of total annual cooling demand. An improvement in system performance of the top 20% of cooling users would achieve energy reductions 300% larger than a similar improvement in randomly selected

group representing 20% of users. The top 10% of residences were responsible for 46% of cumulative annual electric heating energy, and the top 25% were responsible for 74%. The lowest 50% of households contribute just 5% of total annual cooling demand.

Estimated change points average near 65°F for daily average temperatures, but increase slightly at higher temperatures. This could be the result of some combination of better insulation and construction practices, better equipment, or occupant adaptation in warmer climates. The mean of 65°F also suggests that residential cooling could make better use of outside air to cool homes with less energy use. These so-called economizer systems are more common in commercial cooling applications. 1, 2, and 5°F increases in cooling change points are projected to reduce total cooling energy by 10, 19, and 44%, respectively. When the same set-point changes are made by just the top 10% of cooling users, these savings are 8, 17, and 38% of total annual cooling energy. Note that 44% of cooling energy is 6% of all residential electricity sales. Even though the difference of a few degrees is barely detectable by occupants in most circumstances, set-point changes have traditionally relied on significant behavior modification and attention from residents. However, the latest generation of smart and communicating thermostats offers a renewed opportunity to reduce cooling energy with more passive participation by residents.

Smart meter data is also very well suited to studying demand during periods of grid stress. Our work suggests that residential cooling contributes 3-6% of total energy on peak demand days, with roughly half of that contribution coming from just 10% of the population. We can verify that the most intense cooling demand originates in the Central Valley, which is the hottest part of the service territory. However, there is a large amount of variability within narrow temperature bands, which suggests that peak reduction targeting should be based on more than location alone.

Very large potential system and societal benefits are anticipated to be achievable through energy efficiency and demand response programs, and a lot is riding on their outcomes. New methods are needed to make more effective use of the limited time, attention, and money spent on such programs. We have demonstrated that newly available smart meter data can become a powerful tool for understanding customer demand. In particular, semi-physical modeling techniques can recover valuable estimates of unobserved residential characteristics that can be used to better target, plan and evaluate utility programs.

Chapter 4

Estimating occupant activity using smart meter data

4.1 Introduction

In a residential setting, a growing fraction of energy consumption is under the direct control of occupants. As construction and equipment standards tighten and electronics proliferate in homes, the portion of household energy consumption dedicated to the traditionally dominant space conditioning end uses is decreasing. Between 1993 and 2009, heating and cooling energy use on site slipped from 58% to 48% of US residential energy consumption [31]. Over the same time period the appliances, electronics, and lighting grew from 24% to 35% of residential consumption[1]. The portion of residential energy that is not directly controlled is still subject to the needs, lifestyles, and preferences of occupants. Even the type and characteristics of appliances and electronics available in a home are subject to the purchasing preferences of occupants.

It stands to reason that programs designed to save energy through efficiency or to shift the timing of its use through demand response should be informed by an understanding of the preferences, habits, and other human traits that can be collectively categorized as behaviors that influence energy consumption. In practice, such programs are heavily influenced by the engineering assumption that everyone values and pursues efficiency and the economic assumption that everyone carefully tallies up the costs and benefits of every energy related decision before choosing the most economically efficient option [78, 67]. See [100] as an influential example of efficiency potential estimated using these assumptions. Other behavioral insights are present in some programs, but the systematic incorporation of behavioral science into program design, execution, and evaluation is still in its infancy.

Part of the reason for the slow adoption of behavioral insights is that human behavior is messy. People behave differently depending on context. A variety of factors can influence the energy use decisions of a building occupant, including personal preferences, mood, recent experience, impatience, social cues, anchoring, risk aversion, imperfect information, indifference, and inattention [50, 37, 4]. The outcome of a decision might be different in an individual or group context, or based on the forceful advocacy of individuals within a household. With so many interacting and competing influences, it is extremely difficult to predict behaviors from the bottom up. The links between all the contributing factors and observed consumption are too complex to disentangle into causal relationships and attempts to do so tend to be heavily freighted with unverifiable, but consequential, assumptions.

In contrast, this chapter develops and applies an empirical model of occupancy and occupant activity in residences using smart meter data. This approach can be used to categorize the patterns of household consumption without assigning causality or relying on heroic assumptions. This approach has not previously been possible because, prior

[1] In absolute terms, heating and cooling energy end uses shrank by 16% and the electronic, appliances, and lighting end uses grew by 47%. The change in space conditioning was likely influenced by population migration to warmer climates and increasingly mild winter weather in addition to improvements in building shell and equipment performance. The change in electronics energy was likely driven by the proliferation of computing devices for work, entertainment, and communication in homes.

to the widespread deployment of smart meters, the data required to support it have not been available.

The great diversity of energy-related behavioral influences might seem like an impediment to the orderly execution of beneficial policies, but there is also opportunity in the diversity. With more personalized insights into how people use energy, programs can be designed around more realistic models of behaviors, and offers can be more effectively directed toward the people for whom they are most relevant or toward people most likely to respond favorably.

4.2 Background

The problem of how to motivate participation in energy efficiency programs, and the debate over the best ways to estimate the benefits of doing so, have been active for at least four decades, with behavioral scientists contributing to the debate since the beginning [53, 109, 105, 12, 104]. As we look for ways to scale up energy efficiency and demand response's contribution to climate change mitigation efforts and lower their costs, behavioral scientists' perspectives on what motivates program participation is increasingly being taken seriously in policy circles[2]. At the same time, the tools of data science, applied to newly available smart meter data, are beginning to provide more detailed empirical treatments of the full diversity of patterns of energy use that compliment and corroborate the prior work of social scientists [45, 59, 2]. This work is a contribution to this effort, best understood in the context of the work that preceded it.

4.2.1 Prior work

In the early 1990s, Lee Schipper, a pioneering energy researcher, was asked about the disparity between engineered energy efficiency potential and realized improvements. He observed, "Those of us who call ourselves energy analysts have made a mistake. We have analyzed energy. We should have analyzed human behavior" [20]. Much has changed since that interview, but the disparity between potential and realized savings persists, and opinions on the cause remain highly correlated with the academic background and professional affiliation of the opinion holder. [37] provides a thoughtful treatment of the traditional economic view that energy is best understood as a commodity used as an input into the production of specific desirable services. If one assumes that energy consumers follow simple micro-economic models of decision making, it is possible to define optimal levels of consumption in terms of the desired level of service and the cost of that service. From this perspective, deviations from optimal levels of consumption are considered to be the result of real world imperfections — either in markets or people —

[2]See for example the growing popularity of the ACEEE's Behavior Energy and Climate Change conference, the success of behavioral efficiency firm OPower, and the many behavior-based programs documented by [21].

interfering with the optimal outcome. Behavioral economists have introduced prospect theory and concepts like bounded rationality and heuristic decision making as adaptations to neoclassical economics more consistent with observed outcomes, but, the terms used to describe this work, including *behavioral failures, behavioral barriers, bounded rationality, behavioral bias,* and *anomalous behavior,* reveal that the central goal of correcting these imperfections and improving market function remains.

The traditional engineering view of energy is that improvements in system and product design can achieve energy savings while maintaining or improving the service provided. On the leading edge of innovation, efficiency gains will tend to be unaffordable, but the commercialization and adoption of such technologies has the potential to dramatically reduce energy use without sacrificing service or value. New innovations become available as quickly as older ones get adopted, so there are always more efficiency technologies waiting to be adopted. [66] and [38] provide analyses along these lines, concluding that savings of upward of 75% of consumption are achievable and affordable over a couple of decades with known technologies.

A hybrid of the economic and engineering views of efficiency has emerged to support policy discussions about energy efficiency and climate change mitigation. Analyses that marry economic choice models and projections of technology performance and costs into the future are now standard tools for quantifying efficiency and mitigation potential. [15] and [63] are prominent and influential examples of this approach.

Each field assumes a different focus for energy consumers: economists focus on cost optimization, engineers tend to assume that efficiency is its own reward, and policy analysts rely on models of technology adoption. However, the vast majority of consumers don't think about the price or source of their energy very often. Nor do they think about the efficiency of their equipment. Most people don't have access to detailed information on how the systems that power their lives work or the externalities of operating them. Energy is a cheap and reliable service that most people in the developed world have the luxury of not thinking about. Yet it is vitally important. It is on the short list of fundamental requirements for modern life. Whether conspicuous or discrete, energy consumption has become inextricably linked to consumer lifestyles, preferences, and even identities.

This means that different people use energy in different ways, even under very similar externally observable circumstances. For example, [109], as a part of one of the earliest systematic studies of residential energy use, documents variability in heating demand in the same homes with different occupants. Subsequent work has revealed energy use to be the result of a complex set of interactions among technology, economics, culture, preferences, habits, and personalities. Widespread gains in efficiency will depend on the synthesis of ideas from disciplines related to all these factors. [68] reviews the multidisciplinary literature on "human factors of energy use" and argues that social and behavioral aspects of energy use can easily eclipse the traditional tools of engineering and economics in explanatory power over energy outcomes. More recently, [67] documented the behavioral assumptions, often unstated and unexamined, embedded in modern efficiency programs.

Studies of potential savings from behavioral interventions have focused on different aspects of behavior and are not directly comparable. Using survey methods to score energy efficiency behaviors among occupants of 15 homes in the UK, all achieving an EcoHomes excellent rating, [36] was able to explain 51%, 37%, and 11% of heating, electricity, and water usage, respectively. [25] studied behavioral impacts on household energy use through 17 efficiency interventions. They identify *infrequent equipment purchasing* as the behavior most responsible for an estimated mitigation potential of 20% of household emissions. In contrast, [23] focuses on *feedback*, with categories like direct (e.g. real-time in-home displays) and indirect (e.g. additional information on bills), and examines the depth and persistence of savings. She reported savings of 5-15% from direct feedback and 0-10% for indirect feedback — ranges consistent with the overall sense that outcomes are variable and dependent on context and program details. Focusing on *bill-based feedback*, [4] estimate that "non-price mechanisms" developed in the fields of psychology and behavioral economics can be applied to reduce electricity use by 2.7% per household, for a reduction of 37,820 GWh/year, or 12.7MtC in the US. However, in [3], a feedback based program was found to achieve a reduction of 2.0%, which is about 25% less than the savings assumed in their national estimates. [102] demonstrates average household energy savings of 2.7% due to the "Hawthorne Effect," which is the tendency for study participants to perform tasks better when they are reminded that they are being watched. The Hawthorne Effect disappears when participation ends.

With so many competing models for predicting and influencing energy use and so much of the academic discussion of these models focused on establishing which should dominate over the others, there is a need for methods characterizing empirical patterns of energy use to support objective evaluation and performance comparisons of competing models. Successful models of consumption should predict the distribution and diversity of usage observed in practice. Widely distributed smart metering infrastructure has made population scale characterizations possible.

Statistical models and machine learning algorithms are beginning to reveal a great diversity of usage (reflecting the complexity of the factors that influence consumption) and to offer methods for modeling and quantifying that diversity. [73] uses regression models with socio-economic regressors for head of household age, economic class, and the presence of children, along with other regressors related to dwelling characteristics, to explain total energy, peak demand, load factor, and timing of maximum usage as derived from smart meter data. The models' R^2 values were 0.32, 0.33, 0.09, and 0.026, respectively, with socio-economic regressors found to be statistically significant in each model and the weakest fits associated with the timing of usage. Even with access to rarely observed characteristics, much of the variation in demand came from unobserved factors. A literature review in the paper summarized past work on explanatory variable for residential energy use: *"The top four variables, dwelling type, household income, appliance holdings and number of occupants appear frequently in the literature. However, it is important to note that the frequent occurrence of certain variables may also be a consequence of the ease with which data was collected."*

[1] applies clustering techniques to identify recurring patterns in the timing of electricity demand within days. A subset of clusters is interpreted qualitatively based on the patterns in the timing of use. [2] fit patterns of energy use using Hidden Markov Model (HMM) state estimation. They find correlations among hidden state characteristics, variability, and household characteristics. For example, elevated mid-day variability is correlated with unemployment, and high variance in high mean states correlate with the presence of washing machines. [59] apply clustering techniques to the challenge of classifying daily household load shapes. Once classified, similar patterns in demand can be used to segment the population of customers into classes of empirical energy use behaviors for participation in utility programs and services. They ensure a good fit between cluster centers and cluster members by employing *adaptive K-means clustering*, which enforces a fit threshold that, if crossed, splits the offending cluster and re-fits both resulting clusters using the same criteria recursively. The number of resulting clusters is then reduced by combining the most similar clusters using *hierarchical clustering* until the target number is achieved.

4.2.2 Our approach

Prior work has shown that social and demographic characteristics of households and the physical details of building construction and equipment only tend to explain a fraction of the variability of electricity use within and between households. Everything that these variables cannot explain winds up in the model errors. Thus, the imprint of the personality and preferences of the occupants tends to be lost in metrics of model error and the standard errors of model coefficients.

We are interested in what can be learned about utility customers using only their smart meter electricity data. As a proxy for the behaviors of unobserved occupants, this work defines "*occupant activity*" to be based on empirical patterns in smart meter data. Specifically, we classify the highest outliers of a regression model, designed to adapt to seasonal change and follow regular and predictable patterns in energy consumption, as instances of likely occupant activity. We then study the patterns in the timing of such occurrences using clustering techniques inspired by prior work to identify categories of temporal patterns of occupant activity. We compare hour-of-day patterns in occupant activity with hour-of-day temperature response and hour-of-day fixed effects derived from regression model fits. We also show how our occupant activity metric can be used to improve the performance of models that customarily ignore these outliers as a nuisance. Finally, we offer some examples of how the occupant activity metrics can be used to better inform the execution of energy efficiency and demand response programs.

4.3 Model development

This section details the development of the occupant activity metric and the regression model used to derive it. Our goal is to define a model that fits all regular and predictable daily patterns of temperature sensitive and insensitive loads so that the model errors can be interpreted as deviations from these predictable patterns. We hypothesize that occupants making energy use decisions for loads under their direct manual control (e.g. doing a load of laundry, blow drying hair, making a smoothie, or spending an evening watching TV) will produce such deviations and that these deviations will be associated with positive model errors and will not be regularly spaced in time. Instead, their distribution in time will reveal something about the patterns of occupant activity in their homes.

4.3.1 Data set

The data set used for this study consists of records for 30,000 PG&E residential customers, with 10,000 randomly sampled from each of three geographic zones — Coastal, Inland Hills, Central Valley — that roughly cover the territory's climate variability. The populations of the zones are: Coast: 1.1M, Central Valley: 1.5M, and Inland Hills: 1.8M.

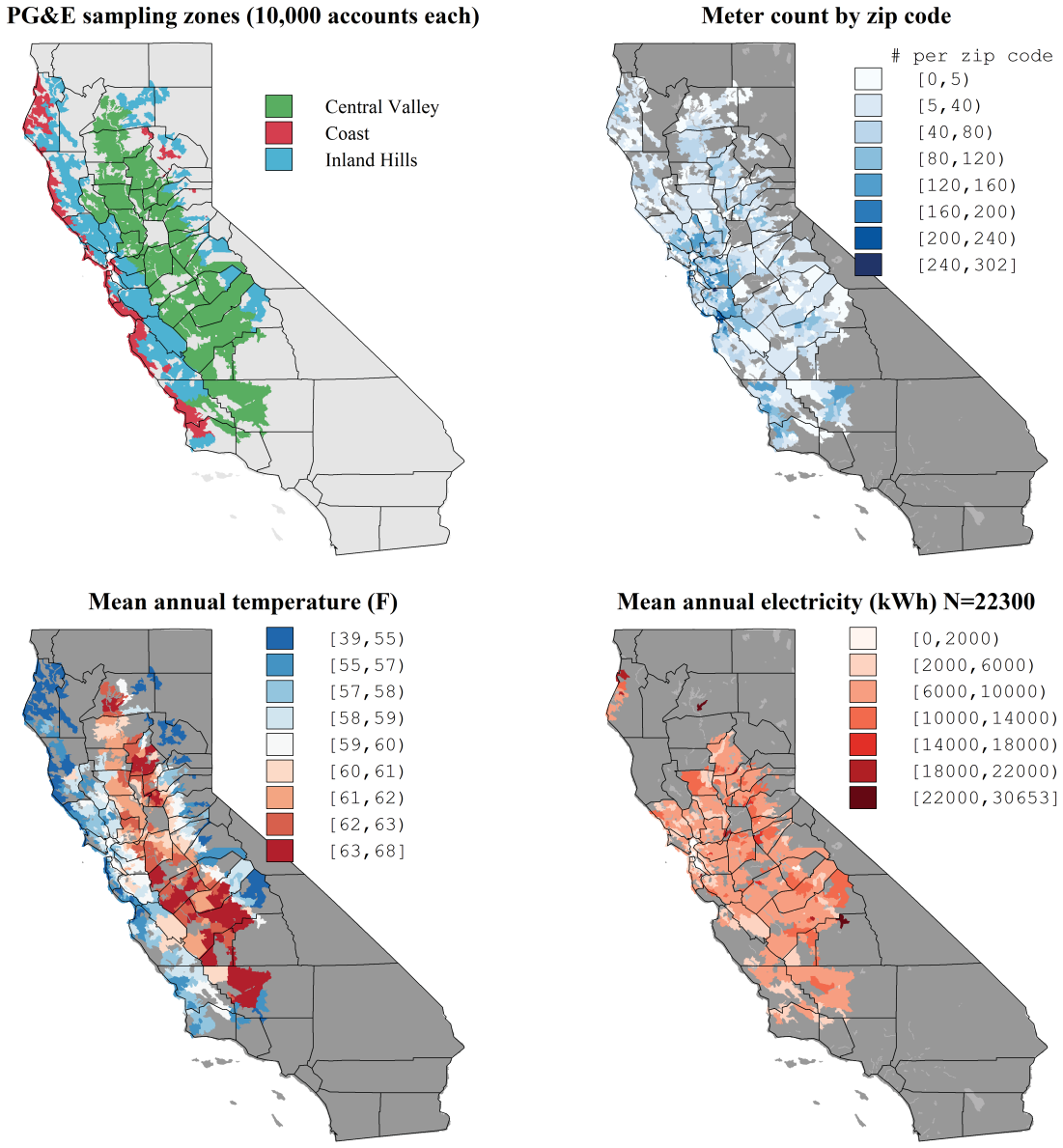


Figure 4.1: Maps of the sampling zones (upper left), the distribution of meters sampled by the data set (upper right), the mean annual temperature by zip code (lower left), and the mean annual electricity consumption in kWh by zip code (lower right).

The coastal climate is mild. The Pacific Ocean moderates temperatures and enforces a winter rainy season, with long, sunny, and warm summers and significant fog coverage. It includes the cities Eureka, San Francisco, and Santa Cruz, and nearly reaches Santa Barbara.

The inland hills see larger diurnal and seasonal temperature swings, but only the

eastern segment of the hills experience freezing conditions and snow routinely during winter. This region includes Napa and Sonoma Counties, San Jose and the Silicon Valley, the East Bay hills, the Santa Cruz mountains, and, farther east parts of the foothills of the Sierra Nevada range.

The Central Valley, on the other hand, is one of the most productive agricultural areas on earth and features the required climate conditions: long, sunny, and hot summers and mild winters. It includes the cities of Chico, Sacramento, Fresno, and Bakersfield.

Once customers were selected, smart meter electricity and natural gas data associated with their account, along with their billing data, zip code, and a subset of customer account records was gathered. This data was provided for study via an open call for research proposals by the Wharton Business School's Customer Analytics Initiative[3].

To support the study of weather impacts on energy consumption, hourly temperature data for each of the 823 zip code was collected from the online weather data aggregator web site Weather Underground, with simultaneous values from 3-5 weather stations averaged, dropping missing observations, to produce an aggregate weather data time series for every zip code.

Figure 4.1 on page 104 presents maps of the PG&E service territory covered by the data set, with the Coastal/Inland Hills/Central Valley geographic zones used to partition the sample of 30,000 customers into 10,000 from each zone (upper left), the number of meters per zip code sampled in the data set (upper right), zip codes colored by their mean annual temperature (lower left), and the zip code values for mean of annual electricity consumption derived from the smart meter data (lower right).

Of the original 30,000 customers, roughly 24,500 meters passed all data validation and cleansing criteria. Validation required more than 180 days of observations, no protracted periods of zero energy consumption, availability of usable matching temperature data from a nearby weather station, and average power demand > 110W.

For every household that passed validation in the sample, we ran the regression model defined in equation (4.1) over a sliding window of data as defined in section 4.3.5 and derived indicators of occupant activity over time according to equation (4.6). These results were saved for analysis along with basic totals and summary statistics characterizing the energy use of each household.

4.3.2 Defining *occupant activity*

If all residences were sub-metered to provide high-resolution data on the use of every circuit in their main panel, it would be fairly simple to separate occupant driven loads from all the others. An inspection of activity on circuits dedicated to plug loads, kitchen appliances, laundry rooms, lighting, etc. would tend to reveal the timing and magnitude of end uses under the direct control of the occupants.

Such comprehensive metering is neither available nor planned for most households,

[3] `http://www.wharton.upenn.edu/wcai/`

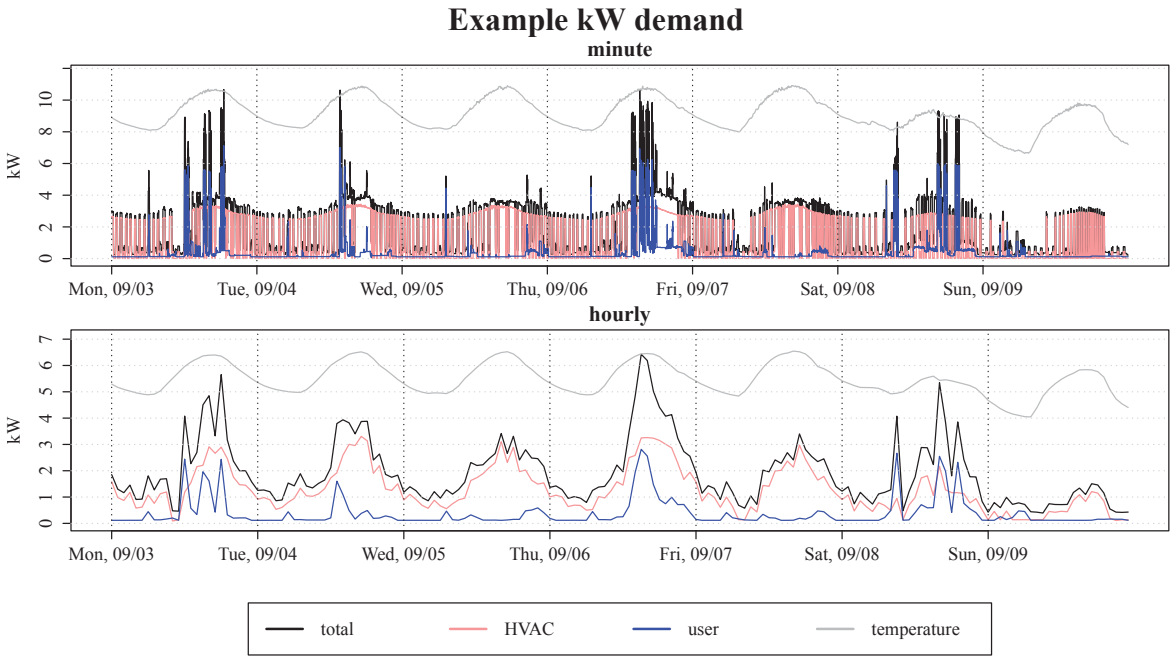


Figure 4.2: Example of residential end use breakdown using 1 minute and hourly data from a real, sub-metered home. Red is HVAC demand, blue is occupant-controlled loads, black is the total demand, and the gray line is re-scaled temperature data for reference. Data is from 2012. Top is 1-minute averaging interval between readings. Bottom is 1-hr averaging for the same data. Data courtesy of Pecan Street Research Institute in Austin, TX.

but there are a small number of instances of homes outfitted with this type of metering in the research literature. [56] and [80] are two prominent examples. Figure 4.2 on page 106 provides a look at sub-metered data from the latter. Visual inspection of the end use breakdowns from the figure shows how the manual control of loads in the minute-by-minute upper plot blue line translates into hourly patterns of the blue line in the lower plot. Occupant controlled loads, known in this case to include a large electric clothes dryer, is additional to the base consumption and air conditioning, which tends to co-vary with outside temperature.

The minute-scale data reveals that both instantaneous magnitude and on/off duty cycle timing of the air conditioning are governed by outside temperature, with each reading strongly dependent on what part of the duty cycle the equipment is in. The hourly averages tend to mask the duty cycles, resulting in cooling loads more directly proportional to temperatures. When the temperature drops low enough overnight, presumably lowering

the internal temperature of the home below its cooling set point[4], there are also periods without any cooling at all. While such patterns are not prominent in this sample of data, it stands to reason that there can be other loads in the home that are governed by deterministic and predictable controls, like timers or daylight sensors that follow predictable schedules. Finally, some non-zero, always-on base loads are likely to be present. These will include refrigeration[5] and electronics that are either always on, like wireless Internet, or whose "off" or standby modes continue to consume energy, as is common with audio and video equipment and plug adapter power supplies.

4.3.3 A generic model of occupant activity

Many of the loads in a home will be predictable with knowledge of outside weather, time of day, day of week, and season of year. If we define the observed power demand at time t as y_t and all the other information we have gathered about the local conditions as the matrix $\mathbf{X}$, with a row of observations for each t, then we can find some generic function $f(\mathbf{X})$ that makes a prediction, $\hat{y}_t$, of what the value of y_t is at time t. If we define the discrepancy between the prediction and actual values as $\varepsilon_t = y_t - \hat{y}_t$, then a generic formulation of this predictive model is $y_t = f(\mathbf{X}) + \varepsilon_t$. For some models, we might also add prior observations of y to the prediction.

The central idea of this work is that simple external observations will fail to predict on-demand direct occupant control of loads, so a significant portion of loads under direct occupant control will be found in the the model errors, ε. Furthermore, because most occupant activity is additional to existing loads, the sign of the errors associated with occupants will tend to be positive, and, in the case of activity that impacts the grid in a meaningful way, fairly large in magnitude.

For a model designed to fits the predictable loads, the pattern of the large, positive, model errors should reflect the underlying pattern of direct occupant control decisions (along with other, less interesting, variation). We can specify an "occupant activity" discrimination function, $O(\varepsilon)$, that returns a time series, o_t, of 0 or 1 values (or continuous 0-1 probability values) that indicate whether y_t was determined to be influenced by occupant activity[6].

The remainder of this section details the development of a specific pair of $f(\mathbf{X})$ and $O(\varepsilon)$. Other formulations are possible. As long as $f(\mathbf{X})$ captures predictable and regular patterns of consumption and $O(\varepsilon)$ is specified with assumptions conservative enough to

[4] A temporary change in the cooling set point is another possibility. A scheduled change would produce a regular pattern over time. Ad-hoc changes by occupants would have neither a regular schedule nor a correlation with specific temperatures.

[5] Refrigerators technically follow duty cycles of their own, which vary in timing and magnitude based on refrigerator usage and defrost cycles. These are often visible in minute scale data, but average out to nearly constant consumption at the hourly scale. Due to the substantial cooling loads of the example home, refrigeration is a modest contributor to the displayed consumption masked by other sources of variability.

[6] In a different formulation, $O(\varepsilon)$ could instead return continuous probabilities from 0 to 1.

isolate significant deviations from predicted operations, the results can be interpreted as observations likely to have been influenced by loads directly controlled by occupants.

4.3.4 Specifying $f(\mathbf{X})$: a predictive model of demand

Regression models are a natural choice for predicting the more deterministic features of residential energy consumption. Indeed, there is a significant literature on regression models applied to the problem of predicting building loads, including [33] and [52]. In their hourly formulations, such models tend to feature terms for time of day, day of week, and seasonal fixed effects, as well as temperature terms that are either linear or non-linear to capture the expected behavior on either side of the internal set point of the residence. For these models, $f(\mathbf{X}) = \sum_{c=1}^{k} \beta_c \mathbf{X}_{t,c}$, where k is the number of columns of $\mathbf{X}$ and row $\mathbf{X}_t$ corresponds to all outside data for time t. In other words, the prediction is a sum of linear terms, with weights, β fit to each column of data.

$$y_t = \sum_{c=1}^{k} \beta_c \mathbf{X}_{t,c} + \varepsilon_t \tag{4.1}$$

Now we need to determine what data we have available to make predictions and how that data can be married with knowledge of energy use in homes to build a predictive model.

It is well understood that outside temperatures drive heating and cooling loads in homes. Conductive heat transfer through walls, roofs, and floors is linearly proportional to the temperature difference across them. A warm home will lose heat to the environment on a cool day, and a cool home will gain heat. The cooling loads can be expected to increase as warm weather gets hotter, and the heating loads can be expected to increase as cool weather gets colder. However, heating and cooling loads can also be expected to have temperature thresholds beyond which they are not required. As determined by thermostat set points, heating demand should be zero above its threshold and cooling demand should be zero below its threshold.

The outside temperature corresponding to the indoor threshold temperature, where heating and cooling loads are at a minimum, is known as a building's balance point. Because the trend in energy demand as a function of temperature is discontinuous across the balance point, it can be modeled as a change point at outside temperature $= CP$. Due to gains from people, appliances, and electronics, and gains from solar radiation absorbed by the roof and transmitted through windows, the value of CP will tend to be cooler than the thermostat set point inside the house, but as set points are changed over time and vary across households, the values of CP should follow along. Using the notation $(x)_+ = max(0, x)$, equation (4.2) describes a regression model with separate fixed-effect coefficients, $\beta_{d,h}$, for every hour of the week, applied to indicator variables $\delta_t^{d,h}$ that are 1 at each specific hour of the week (i.e., 1 for a specific combination of day of week, d, and hour of day, h, and 0 otherwise, for a total of 24×7 columns of 1s and 0s) and two

independent lower and upper temperature slopes, $\beta_{tout_{h,-}}$ and $\beta_{tout_{h,+}}$, across a change point, CP, all three of which are estimated separately for every hour of the day. The change points are determined using a gridded search of change point temperatures in 1°F increments, looking for the temperature that provides the lowest value for the sum of squared residuals[7], and are exogenous to the final model fit.

$$y_t = kW_t = \sum_{d=1}^{7}\sum_{h=1}^{24}\beta_{d,h}\delta_t^{d,h} + \sum_{h=1}^{24}(\beta_{tout_{h,-}}(CP_h - tout_t)_+ + \beta_{tout_{h,+}}(tout_t - CP_h)_+) + \varepsilon_t \tag{4.2}$$

A simpler version of the model with only 24 hour of day coefficients, as described by equation (4.3), will also prove useful when working with less data or where fewer model coefficients are expected to fit well.

$$y_t = kW_t = \sum_{h=1}^{24}\beta_h\delta_t^{h} + \sum_{h=1}^{24}(\beta_{tout_{h,-}}(CP_h - tout_t)_+ + \beta_{tout_{h,+}}(tout_t - CP_h)_+) + \varepsilon_t \tag{4.3}$$

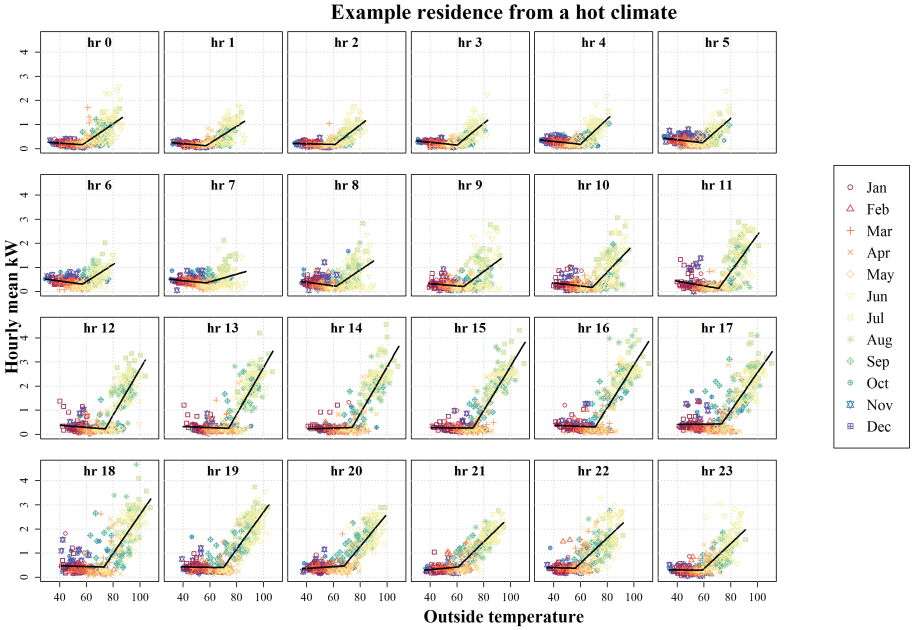


Figure 4.3: Illustration of hourly fits using separate values of CP for each hour of the day, across all months of the year for Home A.

[7]Other methods from the literature on change-point detection could be substituted for this simple grid search minimizing SSR.

Figure 4.3 on page 109 provides a look at the specified regression model applied to a real household's consumption data. The home this data came from, which we will call Home A, is in Bakersfield, CA, and is part of the data set used to perform this study, which is documented in section 4.3.1. The data are broken out into 24 subsets by hour of day, colored by month of the year, and plotted as scatter plots of power demand against outside temperature. The segmented temperature response lines were fit using the model specified in equation (4.3) — fits in the figure use 24 hour-of-day fixed effects and 24 hour-of-day temperature fits, with data and fits segmented into one hour-of-day per plot. The slopes of the temperature response and the change points vary from hour to hour, possibly the result of scheduled changes in the thermostat set point.

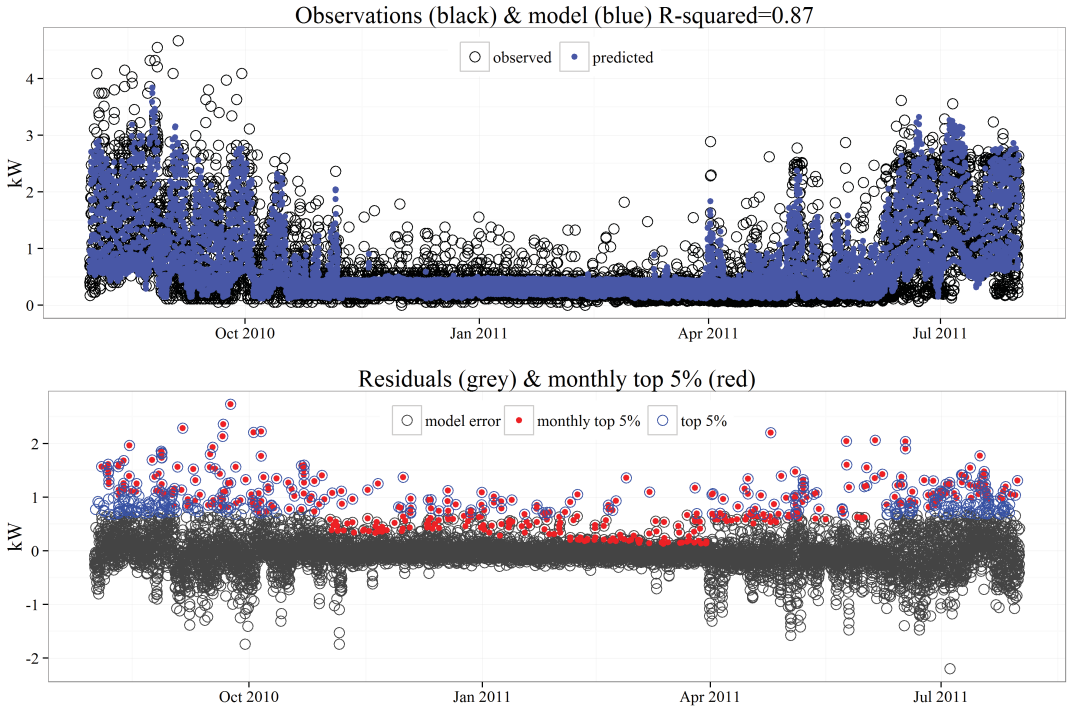


Figure 4.4: Example energy data from Home A with model fit (top) and model residuals (bottom). In the bottom plot, the top 5% of all residuals are in blue rings and the top 5% for each month computed separately are red dots.

The upper plot of Figure 4.4 on page 110 visualizes the hourly demand data from Home A, as a time series. It displays hourly average power demand observations over the course of a full year as open circles and the corresponding model predictions as blue points. The bottom plot displays the model errors over the same time period as open circles.

4.3.4.1 Candidates for $O(\varepsilon)$

The blue rings and red dots of the bottom plot represent candidates rules for $O(\varepsilon)$. The blue rings highlight the top 5% of error values as indicators of occupant activity. The threshold for this group of values is the 95th percentile, which is invariant over time. If this rule defined the behavior of $O(\varepsilon)$, it would flag too much occupant activity in the summer, when energy use and error magnitude are elevated due to increased cooling loads, and not enough in the winter. A simple attempt to fix this bias is illustrated with the red dots. Here the top 5% of errors are highlighted separately for each month. This heuristic guarantees an equal number of activity indicators for every month, but suffers from abrupt transitions between months and tends to force the selection of indicators in winter months. A more rigorous technique for identifying occupant activity is developed in the next section.

4.3.5 Model fits using a sliding window to time

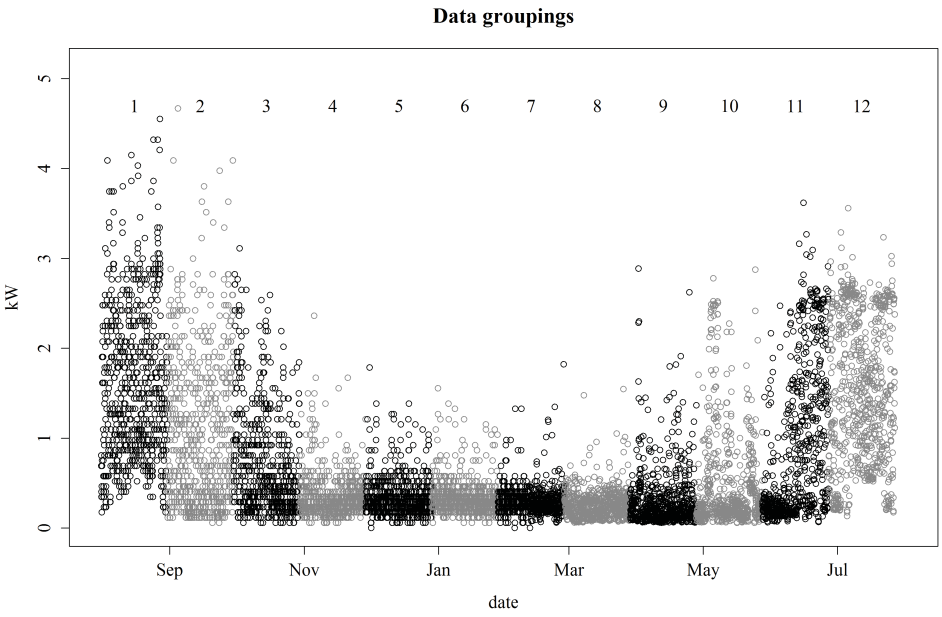


Figure 4.5: Illustration of data groupings used for sliding window regressions, using data from Home A.

In most households, it should be safe to assume that the occupants are present and using their plug loads and large appliances throughout the year. The biggest problem with the candidate rules for estimating the timing of occupant activity from above is that the rules are not capable of adapting to seasonal change or long-term trends in the data.

Patterns in usage from the weeks preceding and following a point should play a large role in determining whether that point should be considered an outlier.

To address this concern, we divide our data, both observed consumption and the associated temperature and date-derived regressors, into 30-day subsets, as illustrated using the time series of points from Home A in Figure 4.5 on page 111, with labels 1 through 12. Then we run several regressions using equation (4.2) repeatedly on a sliding window of points. We have selected a 90-day window of points for each regression and a step of 30 days between each run. Thus the run on sections {1,2,3} would be followed by {2,3,4}, then {3,4,5} and so on. To ensure that every point is modeled three times, we apply special rules at the beginning and end of the data that relax the 90-day rule.

The final sequence of data slices to regress is thus: {1}, {1,2}, {1,2,3} ... {10,11,12}, {11,12}, {12}. We define the set of these data slices to be **S**. After all the models have run, the top 5% of errors from each, notated as $\varepsilon^*_{\{1\}}$,$\varepsilon^*_{\{1,2\}}$, $\varepsilon^*_{\{1,2,3\}}$..., with their timestamps, $\tau^*_{\{1\}}$..., are combined pairwise into a set, $\{\varepsilon^*, \tau^*\}$. Each hour timestamp can occur in this new list from 0 (no models flagged an outlier at that time) to 3 (all three models overlapping at that time flagged it as an outlier) times.

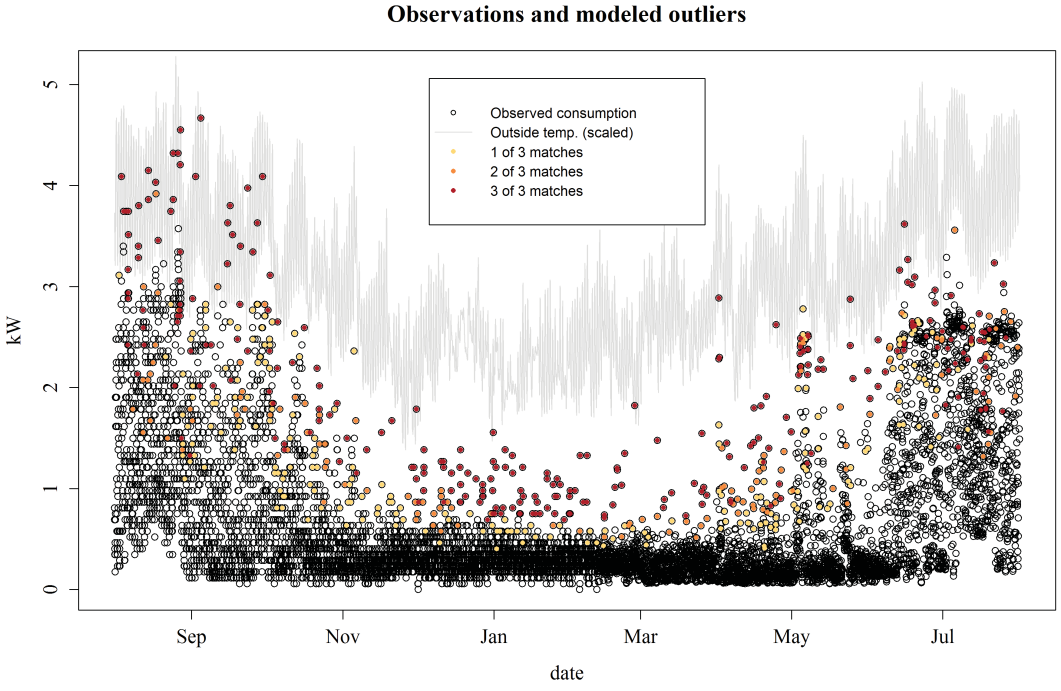


Figure 4.6: Identified outliers in time series, using data from Home A.

Figure 4.6 on page 112 plots the original Home A data, with color coding to indicate which points were matched by this selection rule how many times they matched. As can be verified visually, the matched points are distributed evenly throughout the year, with

significant adaptation to local variance and temperature response. Most of these points are visually out on the high fringe of usage, and, as desired, are likely to indicate some type of occupant-controlled load added on top of already high scheduled and thermostatically controlled loads.

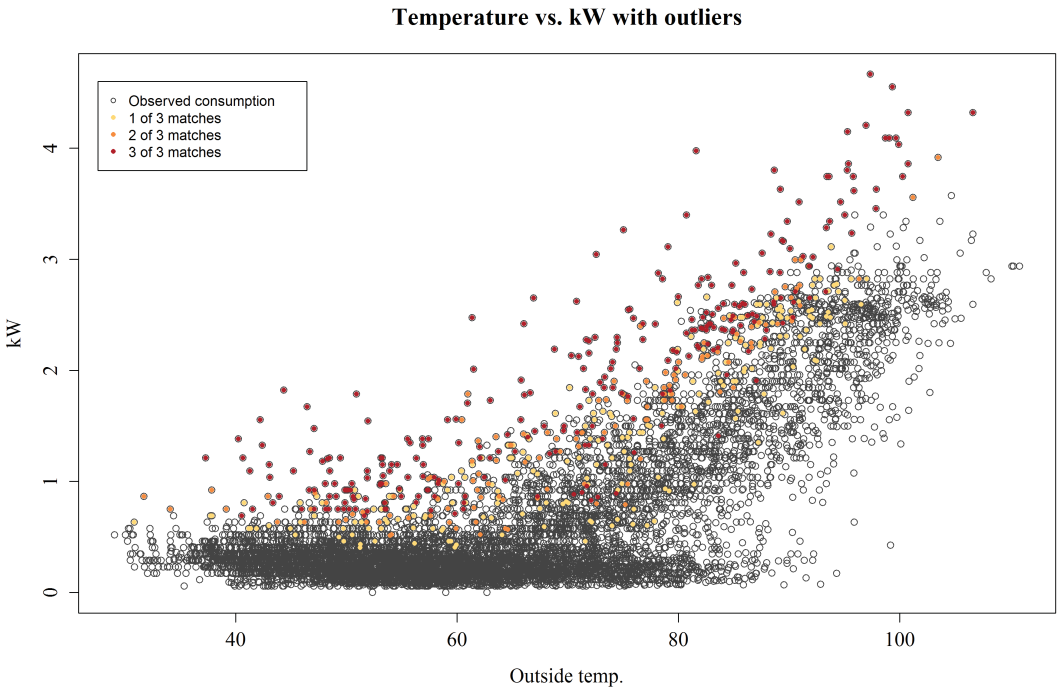


Figure 4.7: Identified outliers in a temperature vs. kW scatter plot, using data from Home A.

Figure 4.7 on page 113 scatters the same demand data against outside temperature. Again we note that the points flagged are consistently part of the high fringe of observed data and are consistently identified across a wide range of temperatures, which includes a change point in thermal response somewhere near 70°F. The highlighted points that are not clearly outside the main cloud of observations are quite likely anomalous for their time of day or day of week. Such points could not be easily found using a simpler heuristic.

4.3.6 A working definition of $O(\varepsilon)$

Now we define our occupant activity indicator function, $O(\varepsilon)$, in terms of a parameter, m, which is the number of overlapping matches required for a given observation to be considered an outlier produced by occupant activity. Defining T as the set of all time for which there are observed values, $\varepsilon^{95\%}$ as the 95th percentile value in ε, and recalling that $\mathbf{S}$ is the set of all sliding window models, we first specify the derivation of ε^* and τ^*.

$$\varepsilon^* = \varepsilon_s^* \,\forall\, s \in \mathbf{S} = \{\varepsilon_s \mid \varepsilon_s > \varepsilon_s^{95\%}\} \,\forall\, s \in \mathbf{S} \tag{4.4}$$

$$\tau^* = \tau_s^* \,\forall\, s \in \mathbf{S} = \{t \mid \varepsilon_{s,t} > \varepsilon_s^{95\%}\} \,\forall\, s \in \mathbf{S} \tag{4.5}$$

$$O(\{e^*, \tau^*\}, m) = \begin{cases} 0 & count(\{\tau^* \mid \tau^* = t\}) < m \\ 1 & count(\{\tau^* \mid \tau^* = t\}) \geq m \end{cases} \forall\, t \in T \tag{4.6}$$

Technically, τ^* is derived from ε^*, so our model of occupant activity from equation (4.6) could be notated simply as $O(\varepsilon^*, m)$.

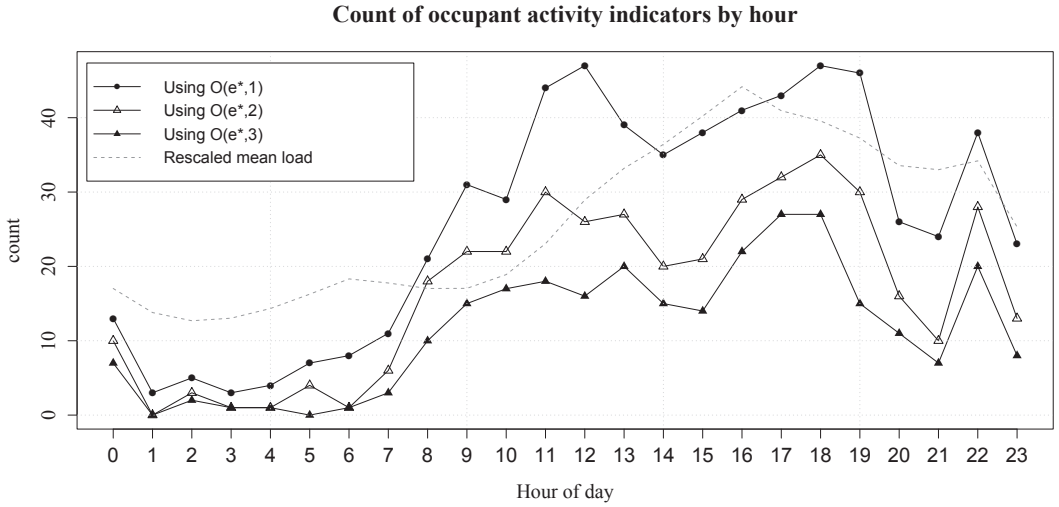


Figure 4.8: Hourly counts of occupant activity indicators for Home A, defined as the points in time identified (from top to bottom) by 1 or more, 2 or more, and all 3 sliding window models as outliers, with a rescaled curve of mean hourly load for comparison.

Now that we have a sequence of indicators for occupant activity for all model observations, we can look for patterns of that activity in time. Figure 4.8 on page 114 depicts a histogram of counts of instances of occupant activity for Home A by hour of the day. The first thing to notice is that despite being derived from model errors, the occupant activity is highly structured in time. The errors of regression models are supposed to be white noise in a normal distribution. However, the pattern of occupant activity by time of day demonstrate that our method is recovering structured data from within the errors. This is an indication that the model is missing at least one regressor with descriptive power and is consistent with the interpretation that the high outliers were generated by unobserved occupant activity. Similar counts can be performed for days of the week, just the outliers within weekends or weekdays, etc. Any of these curves can be normalized by dividing

each value by the sum across all of them. The new, normalized, set of hourly values can be though of as *relative probabilities* of occupant activity occurring. Alternately, the counts can be divided by the total number of observations for each hour, yielding *absolute probabilities* of occupant activity occurring at each hour.

The normalization to relative probabilities will allow comparisons across households and the formation of clusters of similar patterns. The relative probabilities will be notated as $p^i_{hr,h}$ or $p^i_{wday,d}$, where i stands for the i^{th} household, h is the hour of day (1-24), and d is the day of the week (1-7) and the subscripts hr and $wday$ indicate probabilities calculated hourly or daily, respectively. Where context permits, the notation will be simplified.

4.4 Results

With the definition of our regression model and occupant activity indicators worked through, we turn our attention to applying these tools to the sample of data representative of PG&E's service territory. This section presents the results of running the sliding-window regression and occupant activity model against all 24,500 validated residences in the sample. The results presented are based on the most restrictive definition of occupant activity, $O(e^*, 3)$, which requires all three overlapping sliding-window models to identify the same point beyond the 95th percentile threshold for being considered occupant activity.

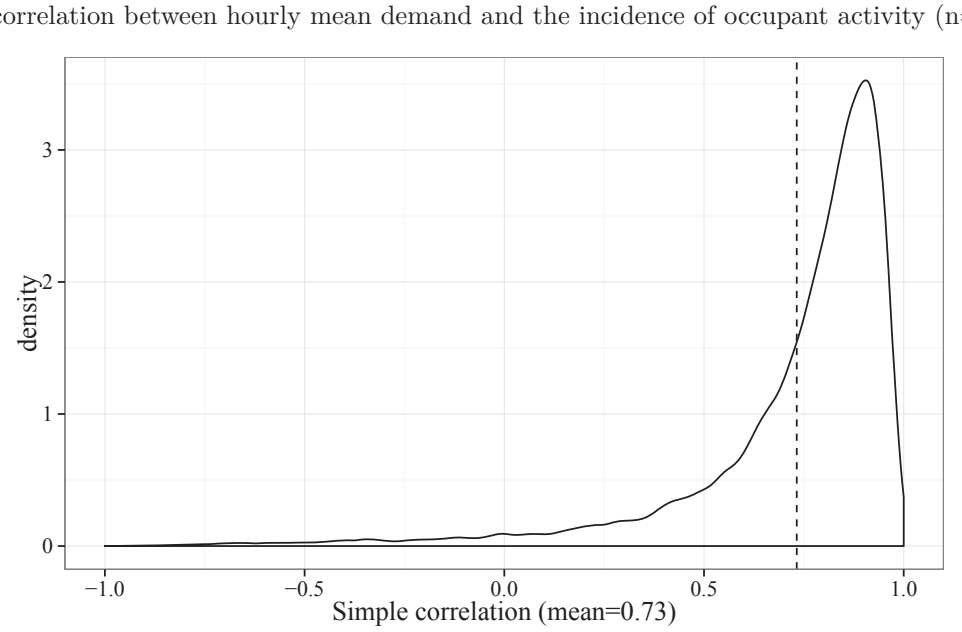


Figure 4.9: Simple correlation between hourly mean kW demand and hourly probability of an occupant activity event. The mean is 0.73.

The outcomes of thousands of model runs can be difficult to examine all at once. To address this, we employ density plots showing the distribution of the full range of values. Figure 4.9 on page 115 depicts a density distribution of the correlation between mean hourly energy usage and the relative probabilities of hourly occupant activity across all 24,500 households in the sample. The mean correlation is 0.73, and it is clear that the overwhelming majority of households have some degree of correlation. One possible explanation for the correlation is that the count of occupant activity indicators is proportional to mean power demand simply because high demand periods will tend to produce larger residuals. However, mean hourly energy use and occupant activity should both be responsive to occupancy. Patterns in mean hourly demand are driven by habitual patterns of occupancy and activity, and the occupant activity indicators are driven by the on-demand consumption of the same occupants. In this interpretation occupants create both difficult-to-predict spikes in usage as well as more predictable central tendencies. Under this interpretation, the degree to which mean hourly demand and occupant activity *do not correlate* is an approximate measure of consumption that may be occurring regardless of occupancy. Section 4.5.2 explores this concept further using the results of the regression models.

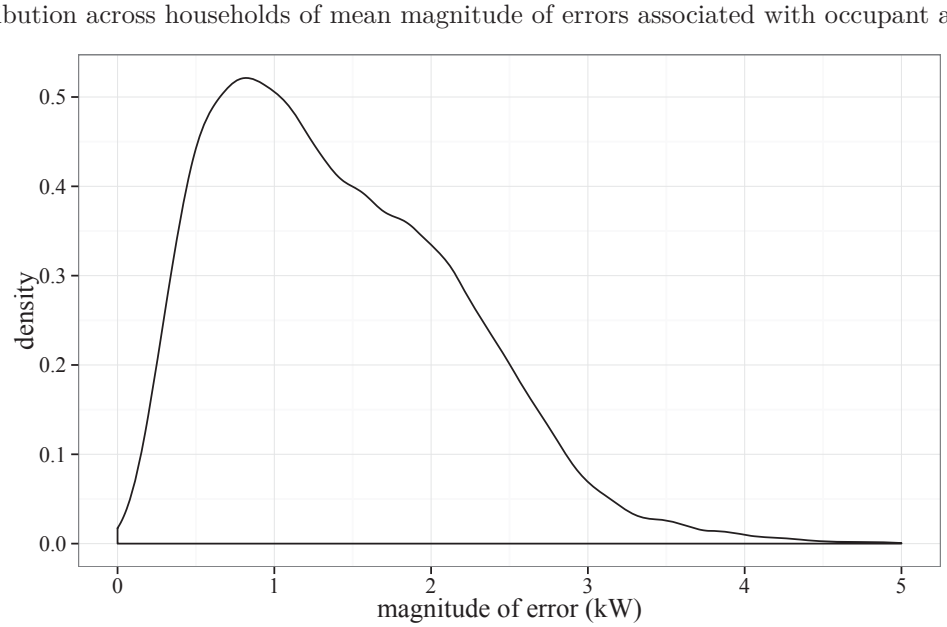


Figure 4.10: Distribution across households of mean magnitude of error associated with occupant activity.

Figure 4.10 on page 116 shows the population-wide distribution of the mean magnitude of the error associated with the occupant activity for each household. Because they are derived from the highest outliers, the magnitude of these errors are quite large for a residential setting. If a model with similar composition to equation (4.2) were used to

forecast household demand, its predictions would be accurate in expectation over time, but its worst errors would often be multiples of the forecasted values. Such errors are of significant concern for demand forecasters, particularly because there are systematic patterns to the timing of such errors, suggesting correlation across households that would prevent such errors from averaging out when aggregated across multiple households.

4.4.1 Clustering the data

With so many residences to work with and given the significant diversity between them, it is useful to classify individual results into groups with similar properties. For many applications, including utility program planning and targeting, the bulk properties of the groups will often be sufficient to guide decision making. This section presents the results of clustering the shapes of relative probabilities of occupant activity using K-means clustering.

To understand K-means clustering, it is useful to think of the relative occupant activity probability results as a matrix of values, with one row per household. In the terminology of clustering, each row will be assigned membership in a specific cluster. The columns contain the relative probabilities of occupant activity for each time period. So for 24 hour of day estimates, the size of the matrix would be $N \times M$, where N is the number of households and M is 24. K-means clustering divides rows of data into K clusters, each of which is defined by its cluster center, which is the column by column average of all of its member's data; each cluster center has dimensions of $1 \times M$. Recalling the notation of p_{hr} for hourly relative occupant activity probabilities and defining the i^{th} cluster center as c^i, the K-means algorithm operates as follows:

1. After the user determines how many clusters are desired, the K initial cluster centers are either selected at random (often using properties of the data for guidance or specified by the user).

2. A distance metric is calculated between each cluster center and each potential member. A typical distance metric will be Euclidean distance, defined for c^i as: $(\sum_{h=1}^{24}(p_{hr,h} - c_h^i)^2)^{\frac{1}{2}}$

3. Cluster membership is assigned for each row based on the shortest distance metric.

4. Cluster centers are re-calculated as the column-by-column averages of their new membership.

5. Steps 2-4 are repeated until cluster membership is not altered by their execution.

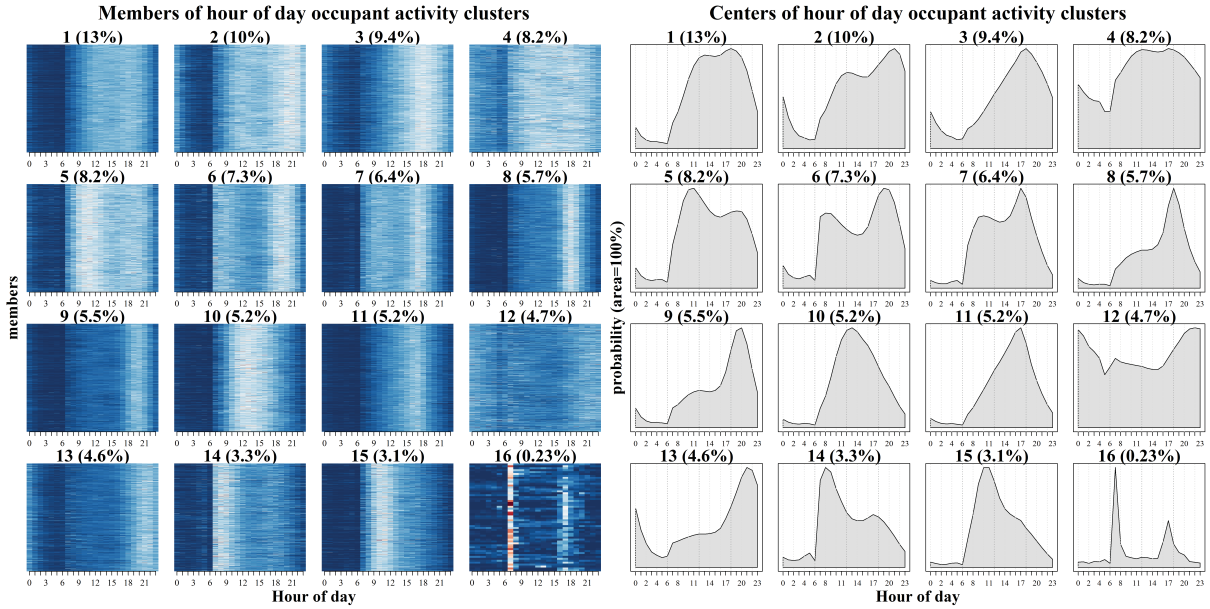


Figure 4.11: Hour-of-day cluster members (left) and centers (right) of relative occupant activity for the entire data sample (n=24,500) with K=16.

Figure 4.11 on page 118 presents the results of K-mean clustering with $K = 16$ run against 24 hourly relative occupant activity probabilities per household. These provide a look at the daily patterns of occupant activity in a form similar to daily load shapes. The clusters are sorted from highest member count to lowest, labeled with their rank and the percentage of households contained by each. The figure on the left visualizes all cluster members using heat maps, whose rows are relative household occupant activity (with 24 hour of day columns) and whose color scales from blue for lowest values to white for mean values and red for highest values. The figure on the right plots the center for each cluster in the same order as on the left. These centers are the column means of all members and can be interpreted as representative curves of occupant activity for each cluster.

As can be verified visually, the timing of occupant activity varies considerably across clusters. For example, 1, 2, 3, 4, 6, 7, 8, 9, 12, and 13 all have significant occupant driven loads in the evening. Clusters 10, 14, and 15 are notable for their lack of evening activity. Several clusters also have quite a bit of activity mid-day, suggesting at least one member of the household is at home mid-day on a regular basis. Most clusters display a marked increase in activity between 6 and 8am. This is, of course, consistent with the rush of morning activity typical of working families and families with kids to get to school. Only 4 (and arguably12) are relatively flat across the full day. These results are highly suggestive of different lifestyles present in the population of utility customers.

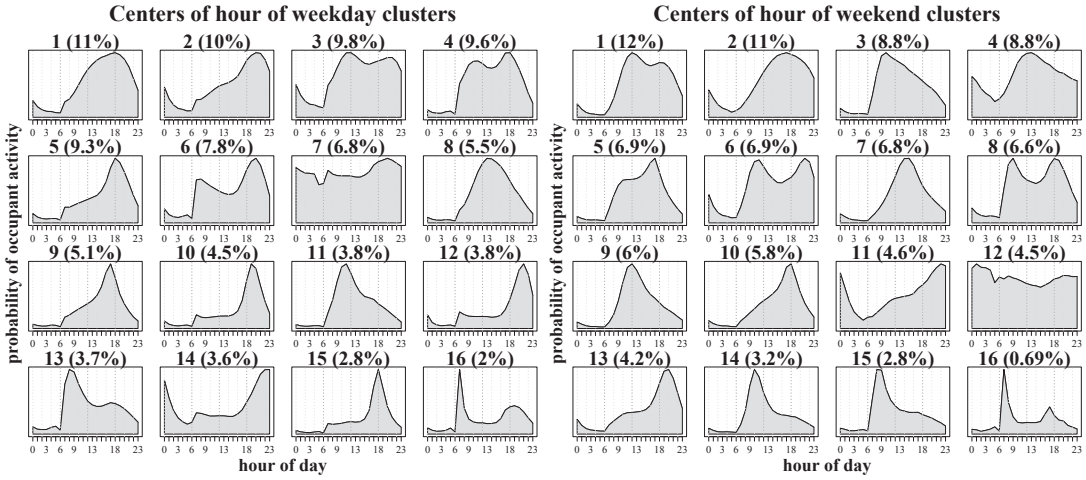


Figure 4.12: K-means weekday (left) vs. weekend (right) cluster centers for hour-of-day relative occupant activity with K=16.

If the cluster centers reflect different patterns of occupant activity consistent with hours of work, it stands to reason that the fits would be different when comparing weekdays and weekends. Figure 4.12 on page 119 provides a comparison between cluster centers for data restricted to weekdays only (left) and weekends only (right). A greater proportion of the weekend centers have greater likelihood of mid-day usage of occupant controlled loads, i.e., the cluster centers tend to increase and peak later in the day, and generally have a less abrupt increase in morning activity. This is consistent with more time at home and fewer scheduling constraints. The weekend cluster membership is also more evenly distributed (in percentages) than the weekdays. This suggests a greater diversity of occupant activity patterns on weekends, again consistent with less constrained schedules.

On a technical note, the minority clusters on the last row of each figure demonstrate a pattern characteristic of K-means clustering. These groups have narrow but prominent spikes in activity. The timing of these spikes is just different enough from some of their neighbors to require separate cluster centers. These clusters, in some sense, relieve the other clusters of odd patterns of occupant activity, improving overall fits.

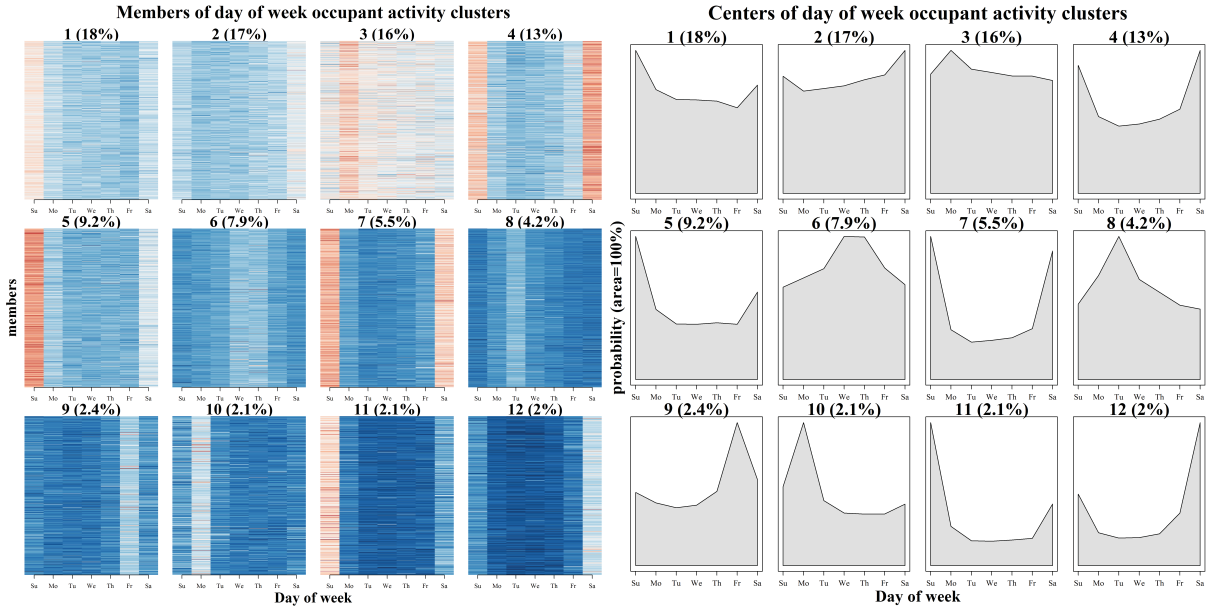


Figure 4.13: Day-of-week relative occupant activity cluster members (left) and centers (right) with K=12. Su=first and Sa=last column.

The differences between weekdays and weekends suggests that occupant activity should show significant variability by day of week. Figure 4.13 on page 120 shows the results of K-means clustering, with $K = 12$, for occupant activity by day of week (i.e. the instances of occupant activity were totaled for each day of week and then normalized into relative probabilities with 7 columns). As we might have expected, several of the clusters show elevated activity on one or both weekend days. This pattern is consistent with occupants spending more time at home — and performing more discretionary energy consuming tasks — on weekend days than weekdays. This effect is not prominent or even present in all households. In clusters 6 and 8, mid-week activity is more pronounced, suggesting the possibility that their members spend the weekends out of the house.

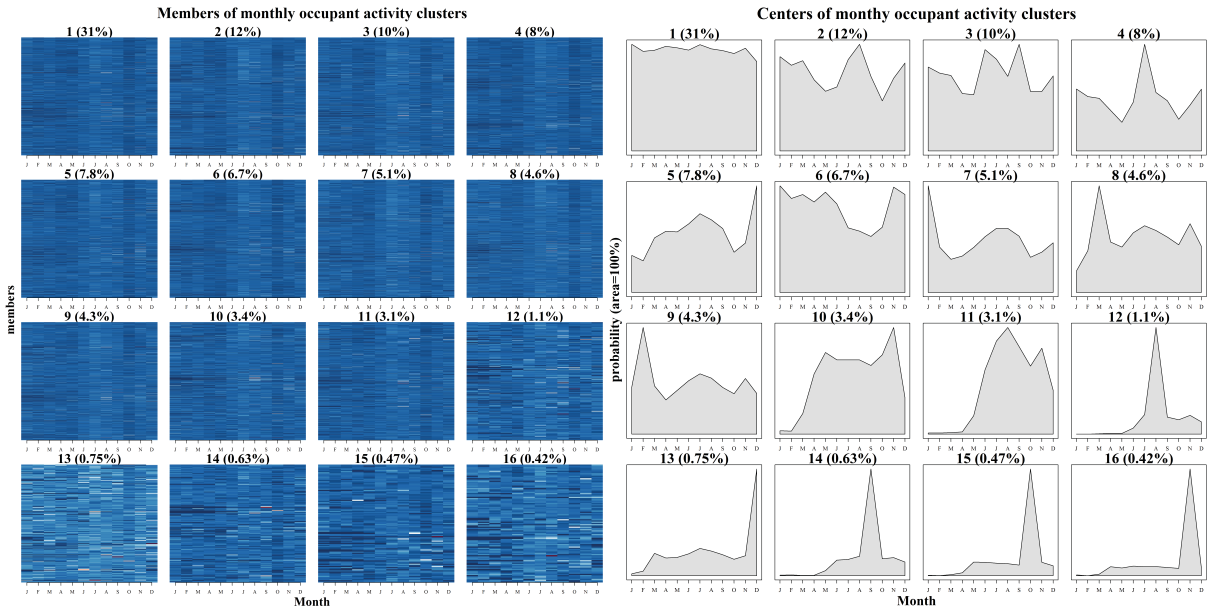


Figure 4.14: Month-of-year relative occupant activity cluster members (left) and centers (right) for K=16. Calculated using 21,260 households with at least a year of data available.

The moving window regression model was designed to address the seasonal bias in the identification of occupant activity that a single annual threshold would carry. It was also designed to avoid forcing a specific quota of fit counts per month. Clusters of monthly relative occupant activity are therefore of interest for a technical reason. They can be used to diagnose how well the sliding-window model has done at distributing the timing of identified occupant activity across seasons, without forcing them. Figure 4.14 on page 121 shows the results of K-means clustering with K=16 for monthly relative occupant activity (i.e., the instances of occupant activity were totaled for each month of the year and then normalized into relative probabilities with 12 columns). Here we restricted the data to samples with at least 8,760 (24×365) hourly observations, eliminating 3,200 households, or 13%, from the analysis. With some exceptions, the clusters are generally well distributed seasonally. In the minority of cases with strong seasonal patterns, we expect that we are observing homes with seasonal patterns of occupancy or structural changes in patterns of electricity use that create a seasonal bias in the model errors.

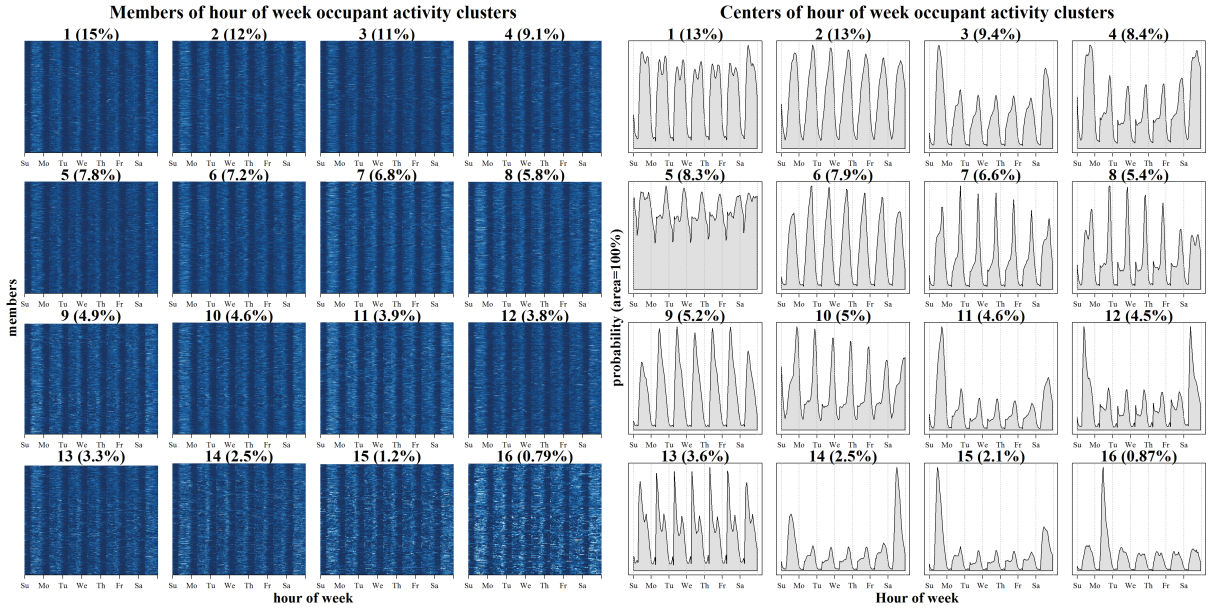


Figure 4.15: Hour-of-week relative occupant activity cluster members (left) and centers (right) with K=16.

For many households, the assumption that all weekday or weekend days will be similar is a poor one. Households are likely to have many different factors that shape the rhythms of their weeks. For this reason, it can be instructive to examine patterns with separate terms for each hour in the week. Figure 4.15 on page 122 shows the results of clustering hour of week relative occupant activity estimates. There are 168 (7×24) columns to each row. In these fits, every day of the weeks is free to have its own load shape. The cluster centers tell the story most clearly. Clusters 4, 11, 12, 14, and 15 all have higher levels of weekend activity than mid-week. Clusters 6, 7, 8, and 9 all have slightly lower peak activity on the weekends, however, 7 and 8 have much fatter peaks on the weekends, preserving the possibility of greater overall weekend activity. Clusters 4, 7, 8, 12 (and arguably others) have qualitatively different shapes for their weekend vs. weekday occupant activity. This is consistent with major shifts in weekday vs. weekend lifestyle. Cluster 13 and the weekends of cluster 12 feature peak levels of occupant activity in the mornings. Most of the rest have evening peaks, with clusters 1, 5, 12, and 13 exhibiting local peaks during both mornings and evenings. Out of all the clustering fits, the hour-of-week fits contain by far the most data per member.

Their 168 degrees of freedom virtually guarantee significant variations among members of clusters and lower overall quality of fits. All of our examples have been based on values of K amenable to discussion of qualitative cluster characteristics. As the cluster membership figures illustrate, the fits were good enough to identify major classes of households with related occupancy activity timing. However, in standard K-means, the distance between cluster members and centers is unrestricted. Some clusters may have very tight fits and

others may have poor fits. The next section addresses this concern by applying a more structured approach.

4.4.2 Adaptive and hierarchical clustering

Because the distance between K-means cluster centers and members is unrestricted, there is often a need to enforce external constraints to ensure good fits. [59] define and apply adaptive K-means to the problem of clustering load shapes with constraints on the maximum acceptable distance between cluster centers and members. The resulting algorithm is available to R users via the 'akmeans' package and documented in [60]. The main idea of adaptive K-means is that instead of specifying a fixed number of clusters, the standard K-means algorithm is run repeatedly with increasing numbers of clusters until no cluster member is farther from its center than some specified measure of distance. After the first run with K_0 centers, all the clusters with members violating the fit threshold are split in two by running K-means with $K = 2$. Any clusters with all members within the threshold are left alone. The splitting is repeated until no members exceed the fit threshold. This produces an arbitrary number of clusters in exchange for a guaranteed quality of fit for all of them, so the number of clusters is determined by the variability in the underlying data. Here the measure of fit, θ for each member, p_{hr} of cluster i, with center c^i, is expressed as follows:

$$\theta = \frac{\sum_{h=1}^{24}(p_{hr,h} - c_h^i)^2}{\sum_{h=1}^{24} {c_h^i}^2} \tag{4.7}$$

In words, the error metric is the sum of hourly squared errors between the cluster member and its center divided by the sum of the square of the cluster center values. A smaller θ ensures a better model fit, and a larger one allows a worse fit.

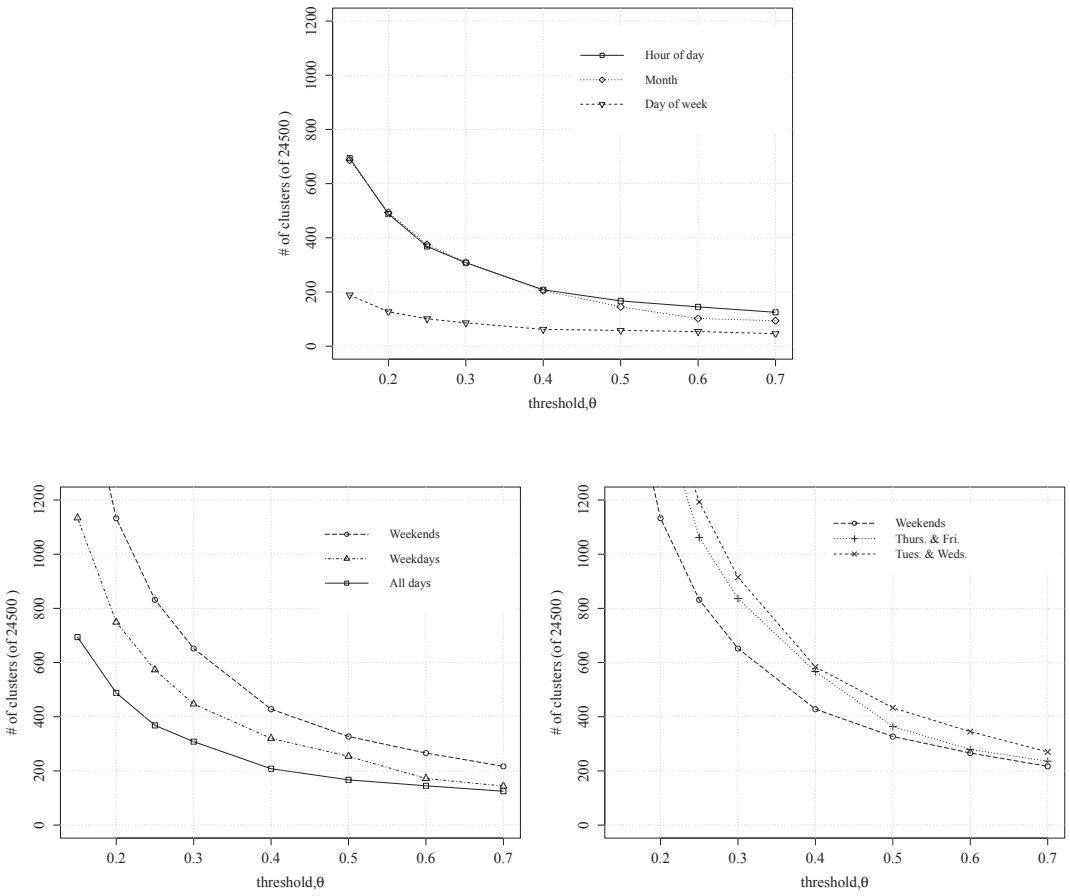


Figure 4.16: Number of clusters required to achieve specific thresholds, θ, of cluster fit for various time scales, e.g. day of week, hour of day, weekend days, etc., of relative occupant activity probabilities.

Given the flexibility of adaptive K-means to add cluster centers to ensure acceptable fits, the operative question becomes how many clusters are required to ensure a given level of fit. Figure 4.16 on page 124 presents three plots with values of θ in $\{0.15, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7\}$ on the x-axis and the count of clusters required to achieve that level of fit on the y-axis, with all scales fixed at 0-1,200. Each line corresponds to a particular set of data to be clustered.

The upper (centered) plot compares the number clusters required for the month-of-year, hour-of-day, and day-of-week relative occupant activity data. Conceptually, the more diverse the patterns in the underlying data are, the more clusters will be required. It stands to reason that the more degrees of freedom in the data, the greater the likelihood of diversity in the data. Since there are 7 days in a week, 12 months in a year, and 24

hours in a day, we might expect the number of adaptive clusters to increase in that order. Consistent with this expectation, the number of clusters for the hour-of-week fits (168 hours per week) exceeds the scale of the plot. Just over 1,200 clusters are needed to achieve a θ of 0.7. However, monthly and hourly values are extremely similar, only separating for larger values of θ. This similarity is partially due to the fact that the overwhelming majority of hours flagged as containing occupant activity fall outside the overnight period. This reduces the number of hours that contribute significantly to the diversity of hour-of-day occupant activity distributions.

The lower left plot compares hour-of-day fits for just weekends, just weekdays, and across all days. Here the degrees of freedom are all the same, yet the number of clusters required to achieve similar goodness of fit are quite different from one another. We might hypothesize that weekends should have greater diversity in their patterns of occupant activity because they are less restricted by work and school. However, we should then logically expect the data from all days, which includes weekends, to be more variable than the data for weekdays only. This is not the case.

The difference in the diversity of occupant activity shapes is largely an artifact of the methods used to derive the relative occupant activity probabilities. Recall that the hourly counts of errors indicating occupant activity are normalized to obtain the relative occupant activity probabilities. This means that the probabilities are, in fact, quantized. Because weekends hours comprise only 2/7 of the data, the difference in probabilities caused by a one-hour difference in counts will be 7/2 as large for weekends as for all days. The weekend shapes will therefore be more likely to have larger differences among members and between cluster members and centers. Similar reasoning applies to the difference between weekdays and all days. This doesn't invalidate the qualitative results discussed above — the clustering results are still internally consistent — but applications requiring tighter control over error magnitudes or comparable cluster fits between clusters derived separately from weekend and weekday data will need to take the effect into account.

The lower right plot compares fits of occupant activity restricted to allow direct comparison between weekend and other day of week data. Cluster fits for weekend data are compared to fits from other pairs of days: Tuesday paired with Wednesday and Thursday paired with Friday. With this control in place, it becomes apparent that pairs of weekdays have a greater diversity of occupant activity than Saturday and Sunday. This result runs counter to the expectation that weekends should be more diverse, but is again subject to a more subtle bias caused by the counting the number of hours with expected occupant activity. If weekends have higher overall counts of occupant activity, they will exhibit less variation than other day-of-week pairs for the same absolute differences in counts. Because of its normalization requirements, K-means clustering is not the right tool to evaluate the degree of absolute variability between categories of data.

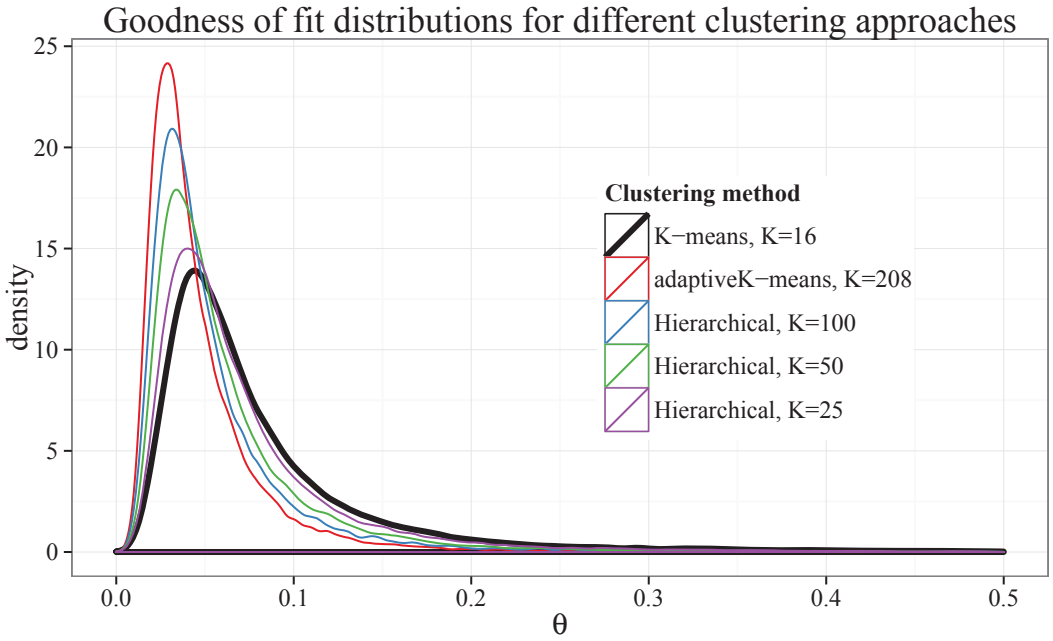


Figure 4.17: Goodness of fit, θ, distributions for various permutations of K-means, adaptive K-means, and hierarchical adaptive K-means clustering.

One major takeaway from the applications of adaptive K-means is that tighter fits of relative occupant activity probabilities across 24,500 residences require hundreds of clusters. However, it is in the nature of adaptive K-means to separate qualitatively similar clusters. Hierarchical clustering is a technique that can be used to quantify the degree of similarity between pairs of cluster centers and rank them from most to least similar. This allows for the systematic merging, also know as pruning, of clusters starting with the most similar. In this manner, the number of cluster can be reduced to a specified count while introducing the smallest possible increase in fit errors. Figure 4.17 on page 126 shows density plots of the error metric θ for hour-of-day data from every household in the sample for standard K-means with K=16 (thick black line), the adaptive K-means fit with θ set to 0.4 resulting in 208 clusters (the tallest, red, distribution), and the results of hierarchical cluster pruning to several intermediate numbers of clusters. As can be verified in the figure, the hierarchical pruning results in an orderly decay in fit.

4.5 Applications

There are several ways to apply the model of occupant activity to improve existing practice in modeling energy use and targeting customers for specific services of offerings. The results of the clustering exercises described above have obvious application to empiri-

cal segmentation based on observed "lifestyle" patterns of empirical activity. This section develops three additional applications of metrics of occupant activity.

4.5.1 Improving model fits

Standard regression models run against household energy data are trying to fit measurements at the meter of many superimposed activities within households. A common application of regression models similar to equation (4.2) is to focus on thermal response in an effort to estimate heating and cooling demand, perhaps with tune-ups or retrofits in mind. However, when such models are given whole home data as their inputs (as opposed to sub-metered HVAC consumption), they suffer from biases due to unobserved activities unrelated to heating and cooling. The outputs of the occupant activity model can be used to remove confounding outliers from the data used to fit thermal models.

By running the model from equation (4.3), with just 24 hour-of-day fixed effects, on all household data (no sliding window) with and without the data points previously identified as influenced by occupant activity, we have the ability to compare model fits with and without the observations flagged as strongly influenced by occupant controlled loads.

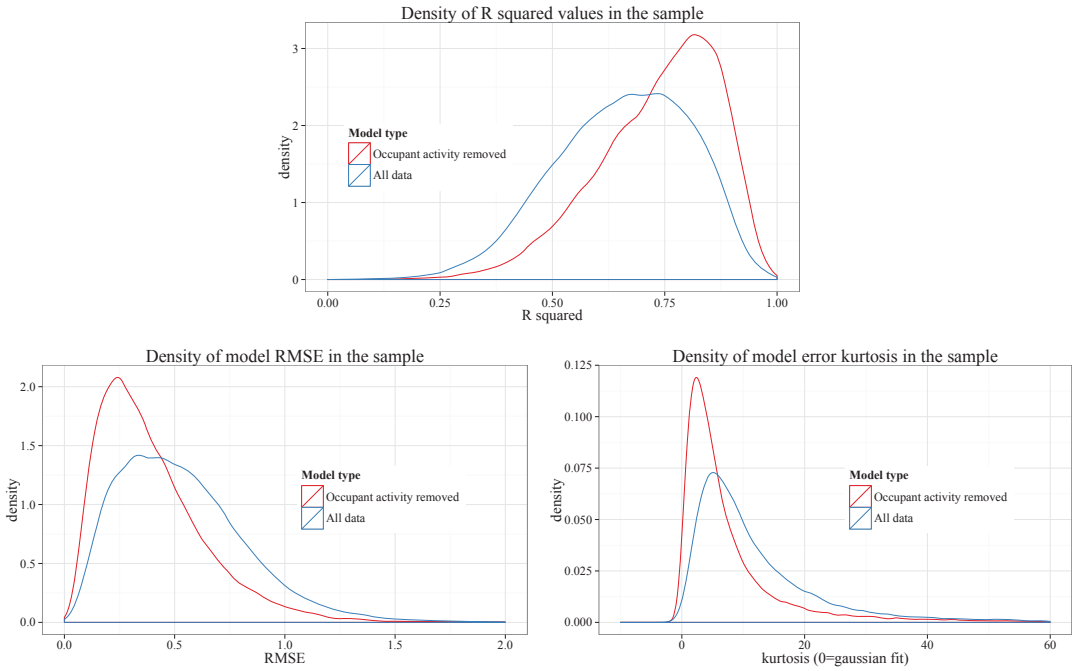


Figure 4.18: Population wide distributions of metrics regression model fit with and without occupant activity points removed.

Figure 4.18 on page 127 presents distributions of three measures of model fit for data with and without occupant activity across the households in our data set. The upper

plot is of modeled R^2 values, which can be interpreted as the fraction of variation in the data explained by the model. The lower left is RMSE, which can be interpreted as the size (in kW) of the typical errors associated with the model, and the lower right is excess kurtosis, which quantifies the shape of the distribution of errors for each model run, with an excess kurtosis of 0 indicating a normal distribution and values greater than zero indicating a sharper peak and fatter tails. Because we have removed outliers from the data, the metrics of model fits are naturally improved. To the extent the model of occupant activity justifies these removals as eliminating data unrelated to thermal response and regular scheduling, the *coefficients* of the models with improved fits should also provide more accurate estimates of thermal response and scheduling.

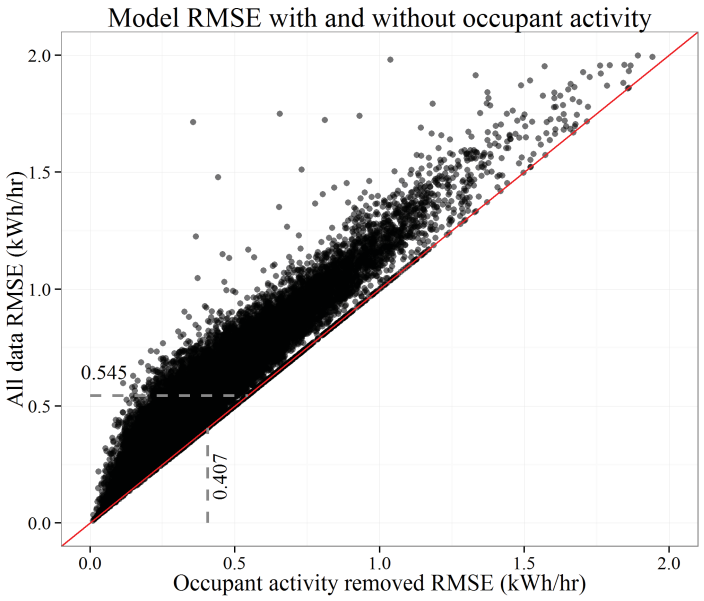


Figure 4.19: Scatter plot of modeled RMSE improvement from removing points associated with occupant activity.

Figure 4.19 on page 128 is a scatter plot of model RMSE. Each household is a point, with the y-axis being RMSE for all data and the x-axis being the RMSE for the data with occupant activity removed. The higher a point is above the diagonal line, the more that household's fit was improved by eliminating observations associated with occupant activity. The degree of improvement is clearly variable from household to household, and this plot has the potential to identify a subset of household particularly impacted by outlying occupant activity.

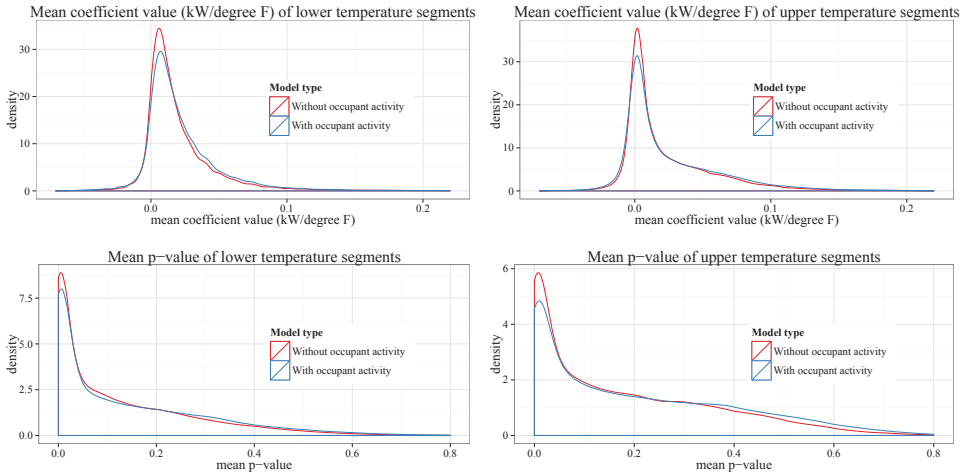


Figure 4.20: Change in mean coefficients and corresponding p-values for models with and without occupant activity data.

The regression model has 24 time-of-day fixed effects and 48 temperature response slopes for lower and upper segments separated by hourly change points. A sense of the magnitude of each of these categories of parameters can be obtained by averaging their values across all 24 hours. Similarly, model confidence in the parameters can be summarized by averaging the corresponding model p-values. The difference in model coefficients and corresponding p-values can be used to assess model sensitivity to outlying data points that are not produced by the dynamics the model captures. These sensitivities are important because regression models that ignore direct occupant control of loads are very common and not generally believed to suffer from systematic biases. Figure 4.20 on page 129 provides distributions of mean hourly lower slope (upper left), mean hourly upper slope (upper right), and the corresponding mean p-values (lower row). As should be expected, the coefficient distributions are tighter, with fewer exaggerated values, and the corresponding p-values are improved overall. However, the overall impact on temperature coefficient estimates is modest.

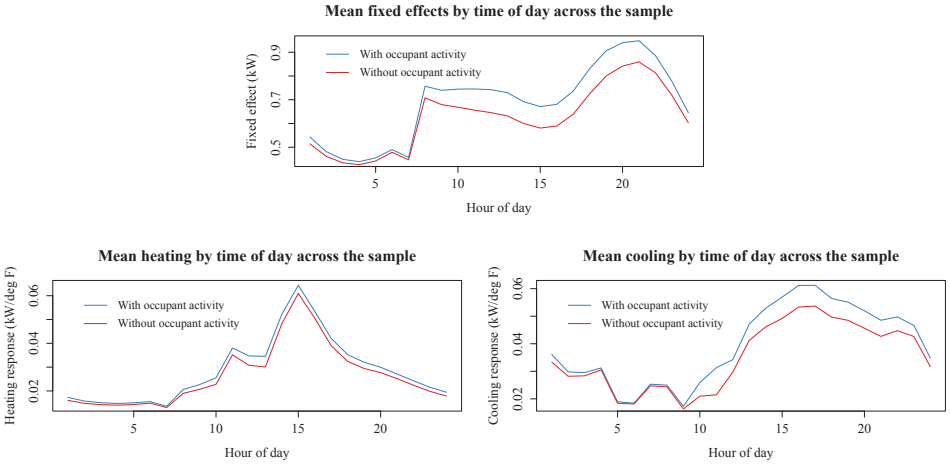


Figure 4.21: Average values of 24 hour of day coefficients for fixed effects, heating and cooling with and without occupant activity.

For many practical applications of building energy regression fits, the model coefficients are interpreted as physically meaningful estimates of energy use. In Figure 4.21 on page 130, we present the impact of removing occupant activity data from the regressed data on hour-of-day coefficients. By comparing the average across all households of hour-of-day coefficients for fixed effects, heating, and cooling for models with and without occupant activity data removed, the three sub-figures reveal the impact occupant activity has on overall model fits. Since higher outliers were removed, the values of the estimates are decreased across the board, but to varying degrees, depending on time of day. The biggest shifts are in the fixed effects and cooling estimates in mid- to late-afternoon. These are times when grid capacity constraints are tend to bind and accurate estimates of sheddable loads are valuable, so a systematic bias in coefficient estimates at these times is especially undesirable. More work will be necessary to determine how robust the change in coefficient values are to reasonable changes in the model of occupant activity.

4.5.2 Identifying mismatched schedules

As a general rule of thumb, energy services, like heat and light, that are provided when no one is home to experience them are wasted. Exceptions to this rule might include leaving air conditioning on for a pet, or lights on for safety or to guide the return home. After accounting for this intentional usage, it is still very likely that at least some portion of energy used while homeowners are out of the house or asleep is wasted. With the understanding that low numbers overnight are probably indications of occupants at home, but sleeping, our occupant activity estimates can be used as an approximate schedules of occupancy. Similarly, model estimates of temperature response can be used as approximate schedules of air conditioning usage, and estimates of fixed effects can

be used as approximate estimates of typical schedules of energy use. By comparing the timing of occupant activity schedules with the timing of hour of day fixed effects and hour of day estimates of temperature response, we obtain a diagnostic indication of whether a home is likely to be wasting energy.

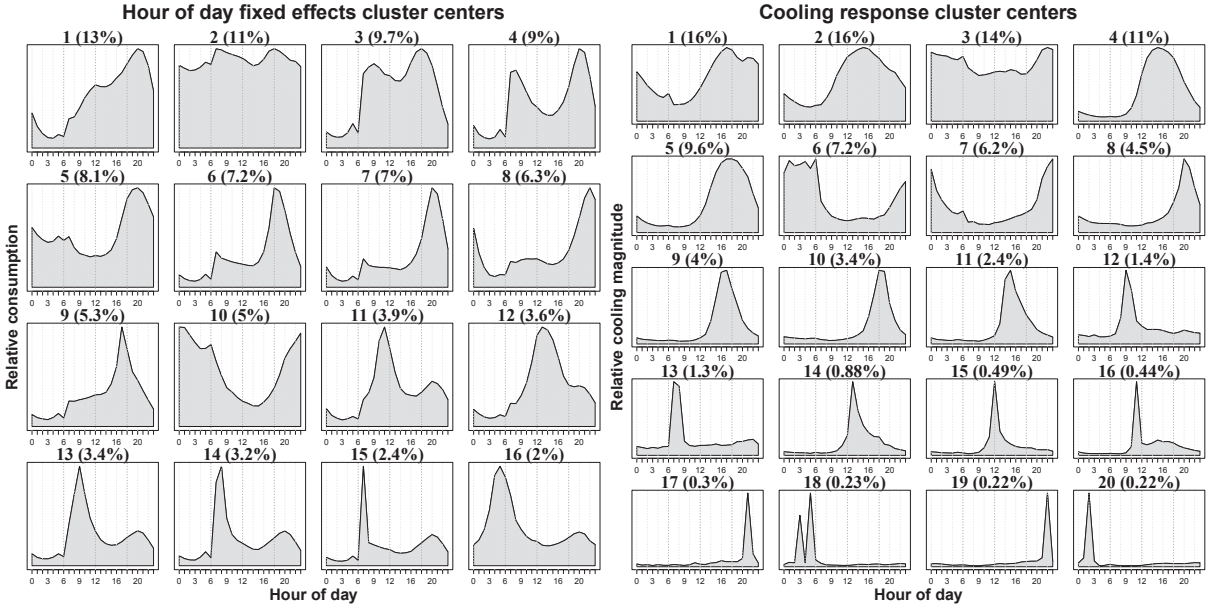


Figure 4.22: Cluster centers for hour-of-day fixed effects (left, n=24,800) and hour-of-day relative cooling intensity (right, n=12,300). To eliminate households without cooling loads, the cooling clusters are restricted to households with an estimate of annual cooling energy greater than 1 kWh per day and are thus fewer in number.

Figure 4.22 on page 131 provides hour of day K-means cluster fits, with K=16, for normalized regression model hour of day fixed effects (left) and cooling intensity (right). It is important to emphasize that K-means used this way requires the cluster members to be normalized so their values sum to 1. For terms with little hourly variation, this can lead to exaggerated swings in normalized values and is likely the source of spikes in the cluster centers.

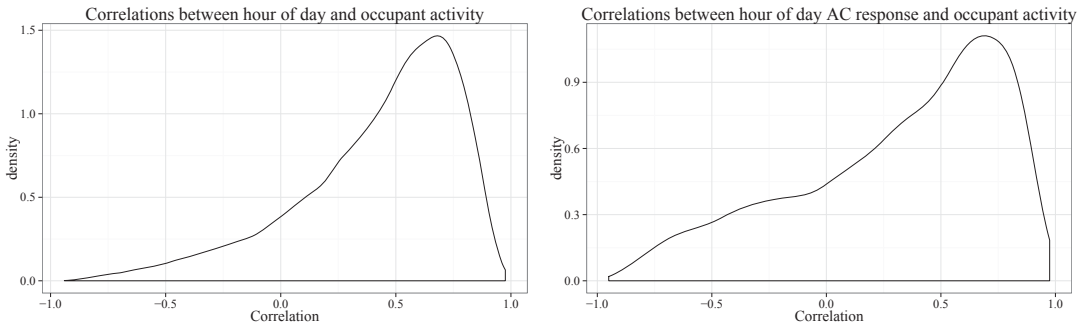


Figure 4.23: Correlation between occupant activity and timing of hour-of-day fixed effects (left, n=24,800) and cooling intensity (right, n=12,300). The cooling intensity results were restricted to households with an estimate of annual cooling energy greater than 1 kWh per day and are thus fewer in number.

As a simple diagnostic of the relationship between the independent schedules, we computed the correlation between estimated hourly occupancy, fixed effects, and air conditioning schedules for each household, with distributions of the results shown in Figure 4.23 on page 132. To exclude spurious results, the air conditioning schedules were restricted to households using, on average, more than 1 kWh/day for cooling. This modest energy requirement is passed by just 12,300 households, out of 24,800, indicating that roughly half of our sample does not utilize significant amounts of air conditioning, a result consistent with northern California's mild climate.

In simple terms, we should expect that a higher correlation between occupant activity and the other estimated schedules indicates a good fit between when occupants are home and active and the more predictable consumption of energy. By the same reasoning, lower correlations can be expected to point to controls that are out of step with the occupants.

Both the hour-of-day fixed effects and air conditioning response timing tend to correlate with occupant activity. However, the air conditioning distribution has a fatter tail in the direction of worse correlation. This is consistent with occupants directly controlling most of their home's energy, but with some thermostats operating according to schedules that do not match occupant activity. Households with low or negative correlation between their cooling loads and occupant activity are strong candidates for closer inspection for inadequate controls over their cooling.

4.5.3 Identifying passive and active occupants

The probabilities of occupant activity offer some predictive power over the timing of the largest deviations from regularly scheduled energy use as fit by a regression model. This predictive power can be useful in determining which households to engage in demand response programs. If we are intervening with automated controls, we might choose to focus on households without much occupant activity to ensure that the controls are less

likely to be overridden and improve the fidelity of subsequent measurements of participation. If we are looking for people to educate about the benefits of shifting loads, like doing laundry, to off-peak periods, we might target households more likely to experience large power swings driven by occupant activity during critical periods.

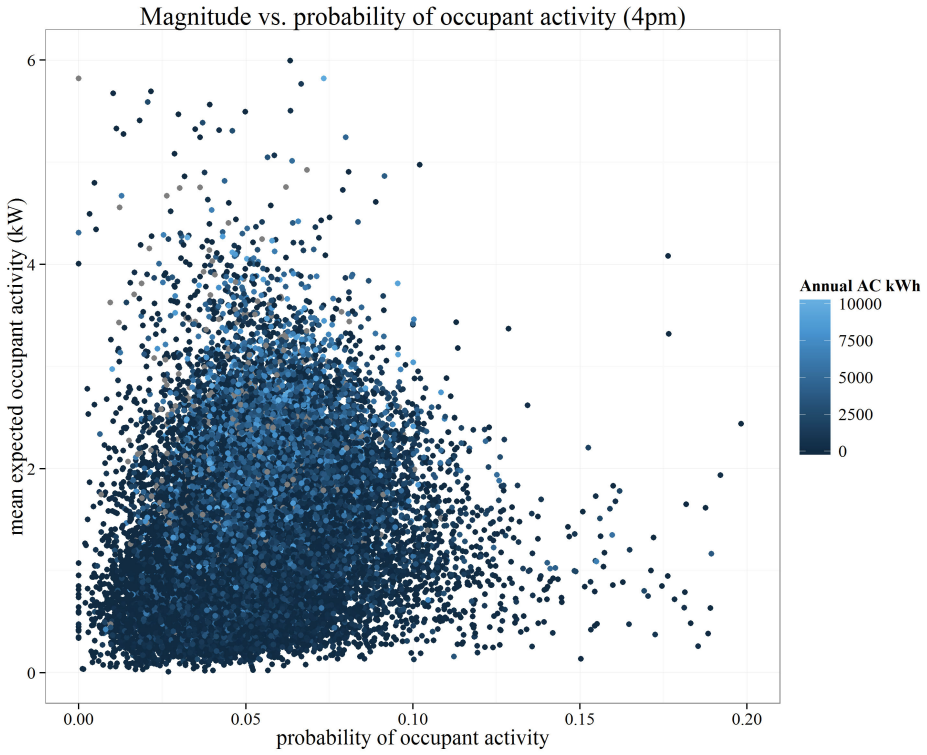


Figure 4.24: Mean magnitude of occupant activity vs. probability of occupant activity at 4pm for all households in the data set.

Figure 4.24 on page 133 provides a view of the model results relevant to demand response. It is a scatter plot of the mean power associated with occupant activity against the absolute probability of that activity at 4pm, defined as the number of hours of occupant activity for 4pm divided by the total number of observations at 4pm. The points are colored by the estimate of the total annual cooling energy for each household. This plot allows selection of households with *active and unpredictable occupants* characterized by a high probability of high power occupant activity events. Alternately, households with more *passive occupants and high cooling loads* might be targeted by programs that make automated changes to cooling set points.

4.6 Discussion

We have built an empirical model to estimate when occupants decide to directly operate electric end uses. Because occupants have to be home to make those decisions, our model can also say something about the timing of occupancy itself, and because the model is based solely on smart meter data, it can be applied wherever such data is gathered.

Our model of occupant activity is predicated on a set of assumptions that cannot be fully verified with the data at hand[8] and we do not contest that other models capable of estimating occupant activity could improve upon our initial pass, or that reasonable changes to our approach would produce somewhat different results. Indeed, we anticipate that improvements will be possible, especially using models trained against data sets with ground truth information about occupants and sub-metered loads. However, we submit that the framework for analyzing and interpreting the resulting estimates of occupant activity will prove robust to different specifications of the underlying processes. The qualitative results we achieve are strongly suggestive of meaningful real-world patterns and of the integrity of the framework used to cluster and analyze the estimates of occupant activity probabilities.

The model we developed relies on the observation that occupant-controlled loads tend to be additional to scheduled and automated loads and that, for a well-specified model of regularly recurring consumption, such activity will lead to large model errors with positive signs. Of course, not all occupant activity consumes a lot of energy, so our model is using the high-energy events — making toast or coffee, doing laundry, microwaving leftovers, watching a big screen TV, etc. — as proxies for the others. We assume, in other words, that the patterns of the highest model outliers over time will tend to reveal general patterns in direct control of loads exercised by occupants. From a demand management perspective, the larger loads are of the greatest interest.

The higher the outliers are, the more confident we become that occupant activity was involved, so there is a tradeoff between the number of events detected and their expected quality. For the applications discussed in this paper, false negatives (i.e., missing events) are preferable to false positives (i.e., mislabeling normal activity as an event). We selected the top 5% of errors, independently identified by our model fit to three different slices of data, as conservative indicators of occupant activity.

The relevant applications of the model outputs — studying categories of occupant "lifestyle" for information relevant to efficiency and demand response programs, targeting programs at customers most likely to respond and benefit, and estimating population-wide potentials for behavioral interventions — are also robust to significant uncertainties.

[8]The best way to validate this type of model would be to assemble a data set of sub-metered homes with loads directly controlled by occupants measured separately from the rest. Such a data set would support the use of improved supervised learning algorithms and the development of metrics of the precision and accuracy of modeled estimates of occupant controlled loads.

4.7 Conclusions

We have developed a model of occupant activity for residential electricity customers using smart meter data and have taken patterns in the timing of occupant activity as evidence of underlying lifestyles of occupants. Our methods reveal patterns of activity consistent with prior understanding of relative activity of occupants between night and day, weekends and weekdays, and seasons of the year. However, they also reveal significant diversity of patterns within the population consistent with the presence of night owls, weekend warriors, commuters, and stay at home parents. This type of lifestyle segmentation is an important goal of efficiency program design and marketing.

We have quantified the extent to which data outliers associated with occupant activity can confound regression models that use whole home meter readings and outside temperature data to model heating and cooling responses. Such models tend to have weak controls against the very real possibility that unobserved occupant activity is an important omitted variable. To the extent that this is the case, those models can suffer from estimation biases and mis-calibrated model error estimates.

We have shown how occupant activity estimates can be used to identify both active and passive occupants. Occupants of both types have useful, but differentiated, roles to play in the context of demand response programs. Active occupants could be given information about the benefits of changing the timing of their usage whereas passive occupants could be expected not to override automated demand response technologies.

We have also shown that the pattern of occupant activity tends to correlate with estimated cooling schedules and with recurring usage patterns. In homes exhibiting weak correlation, one plausible explanation is that control or scheduling problems have placed the home on a different schedule than its occupants. Such oversights are easily fixed once identified.

Our future work will prioritize the use of sub-metered data sets to validate model results and characterize their strengths and weaknesses. When improvements to the estimation of occupant activity are made, attention should be paid to whether and how they alter the qualitative results.

Efficiency and demand response programs designed with insights into occupant behavior represent significant potential for improvement over current best practices. We have shown that useful behavioral information can be extracted from smart meter data. Such insights should be tested and improved upon in academia but also applied in practice, because efficiency research was never intended to be a theoretical discipline.

Chapter 5

Public-interest uses of smart meter data

5.1 Introduction

Public benefits from the use of smart meter data could include improved utility operations at lower cost; more effective efficiency and demand response program planning, implementation, and evaluation; better targeting of rate plans and program incentives; more informed customers; and better targeting of private sector goods and services related to distributed energy, efficiency, and demand response. However, many members of the public and public advocates are concerned about the privacy implications of the new metering technology. Conventional wisdom dictates that these concerns can be addressed by removing identifying information like names, addresses, and account numbers from the data, but recent advances in statistical methods of *re-identification* have demonstrated that such precautions can be inadequate. Correlations between protected data and public information like voting records and online activity can sometimes be strong enough to recover the identity of the person associated with protected data.

This paper is about achieving the potential benefits of meter data analysis while minimizing the risk of disclosing information customers consider private. The central insights of this work are that the development of algorithms to extract useful information from customer data can be done without direct access to sensitive data and the outputs of such algorithms can be designed to meet well specified privacy criteria. The depth of access to customer data can be determined based on the degree to which each entity requesting access can be expected to work in the public interest. For example, academic researchers, with the appropriate contractual protections and data security practices, might be granted access to detailed usage data for the purpose of developing new methods of analysis while product marketers might only be allowed access to spatially aggregated information on consumption patterns relevant to their products or the data of customers that choose to share their data for commercial purposes.

One problem with the evaluation of privacy risks is that the information people consider private can be highly subjective. The speculative nature of hypothetical scenarios can exaggerate the viability and significance of disclosure risks and thereby distract from practical solutions. To steer the discussion of potential benefits and harms toward concrete applications, we provide examples of analyses performed using customer data from Pacific Gas and Electric (PG&E), a public utility that serves 4.5M residential customers in northern and central California, and discuss use cases relevant to deliberations over applications of customer meter data. California's deployment of 11.4M smart meters is nearly complete and large volumes of meter data have been flowing for several years. While details will vary from one jurisdiction to the next, the lessons learned from California's experience are relevant to deployments elsewhere.

Section 5.2 provides background information on the motivation behind California's investor owned utility Smart Meter implementations and discusses the logic of the associated cost and benefit calculations. Section 5.3 develops a list of benefits that could theoretically be achieved through the use of smart meter data and a list of potential harms that could offset those benefits. Section 5.4 discusses the literature on unintended

data disclosures and discusses legal, technical, and procedural strategies for minimizing the potential harm associated with beneficial data disclosures, including our suggestion, *delegated analysis*. Section 5.5 develops concrete examples meter data analysis and discusses potential applications and data usage of each. Sections 5.6 and 5.7 conclude with a discussion of the options available to policy makers and regulators for making use of meter data in the public interest while minimizing potential harm to customers.

5.2 Background: The California experience

In proceedings concluding between 2006 and 2008, the California Public Utilities Commission (CPUC) authorized the deployment of smart meter infrastructure separately for each of the three public utilities that serve most residents of California [1]. As a part of the proceedings, the CPUC found that the new meters would give customers greater control over their energy use and that the benefits of the new infrastructure would outweigh its substantial costs[2]. However, a large portion of the costs are being incurred up front, with the benefits anticipated to accrue over the following 20 years or more. For example, according to [43], Southern California Edison's budget was $1.6B for installation of 5.3M meters, with another $1.6B in projected operations and maintenance costs over 20 years. With financing costs added in, SCE's ratepayers are expected to pay about $5B in smart meter-related costs over 24 years. The projected benefits, including a $1.5B reduction in meter reading costs over 20 years and $3B in avoided infrastructure costs from demand response programs, total over $7B for the same time period. Assuming everything goes according to plan, the benefits from the smart metering program will offset its costs over the lifetime of the meters.

As it happens, the deployments have not been going as planned. Meter vendors have not always been able to keep up with demand, and some deployed meters have suffered from performance problems. In addition, utilities have faced unanticipated protests, civil disobedience, and even lawsuits from ratepayers over privacy, billing irregularities, public health, and other concerns (real and imagined) rooted in distrust of the motivations and information accompanying meter deployments. [42] provides an analysis of the social factors contributing to persistent public health concerns about the radios used to transmit meter data. Acknowledging that their rollout was probably executed with too little customer outreach and education, the Chief Customer Officer for PG&E stated that, "I don't believe we did a good job of seeing the world through the lens of the customer" [3].

[1] Separate press releases for each announcement in chronological order can be found for PG&E (7/20/2006), SDGE (4/12/2007), and SCE (9/18/2008).

[2] According to the CPUC's approval press releases (see utility specific press releases from 2006-2008 at http://www.cpuc.ca.gov/PUC/energy/Demand+Response/benefits.htm), the deployment budget was roughly $4B for 11.4M meters, about $350 per meter. PG&E's deployment costs for 5.2M meters has been budgeted at $2.2B (originally approved at $1.7B). San Diego Gas and Electric's deployment costs $0.6B for 1.4M meters. Southern California Edison's deployment of 5.3M meters has a budget of $1.6B.

[3] See coverage of PG&E's press conference statement on 5/10/2010 accompanying the release of a

Many of these issues can be classified as growing pains, but the unexpectedly hostile reception of smart meters required the CPUC to intervene through a decision allowing customers to opt out of the smart meter program and has delayed the schedule and anticipated coverage of the final deployments. These modifications have, in turn, increased the deployment budget and undermined a portion of the anticipated automated meter reading savings.

Still, as of this writing, the deployments are nearly complete throughout the state and interval meter data is available for most customers. The programs have incurred a substantial fraction of their planned deployment costs, but are just beginning to accrue public benefits. This situation has ratepayer advocates like [43] tallying the money already spent against modified savings estimates and questioning whether the incurred costs will actually be offset by future benefits.

Fortunately, there is a growing awareness that the data delivered by smart meter infrastructure can be used to derive value across a much wider range of applications than were considered by the analysis authorizing the deployment. For example, meter data used to improve energy efficiency program planning, implementation and evaluation would ensure the better use of over $1B in efficiency program spending a year[4]. When meter data is used to make consumption visible, better educate customers, and reveal savings opportunities, it can help to drive market demand for energy efficiency, demand response, and distributed renewable energy. When meter data reveals recognizable patterns of consumption that hold the signature of specific categories of consumption and waste, it can help private businesses refine and better target the energy services and products they develop. Taken all together, these additional public benefits have the potential to substantially exceed the returns currently expected of the meter infrastructure — and these benefits could provide both cost savings and support for the market transformation needed to implement California's Long-Term Energy Efficiency Strategic Plan and reach greenhouse gas emissions targets.

However, the data involved in achieving these potential benefits can also reveal potentially sensitive information about utility customers. As custodians of the customer's data, both the utilities and their regulators are entrusted with ensuring the privacy and security of customer data. Wholesale public disclosure of detailed customer data would no doubt lead to many creative uses, but there could be no expectation that all of those uses will be in the best interests of customers. The challenge before utilities and their regulators is to determine how to encourage the use of data in support of innovation in the public interest while minimizing the potential for misuse of the data.

Recognizing both the opportunity and the risks, the CPUC has begun a series of discussions focused on the idea of a repository of customer data that would be managed for use in public interest as originally proposed in [61]. Much of the discussion to date

document reviewing the rollout process.

[4]According to [8], California's 2010 expenditure was $1.1B, or 24% of the national total. The same report projects national annual energy efficiency spending of $9.5B (middle estimate) to $15.6B (high estimate) by 2025.

has focused on recently revealed weaknesses in standard prescriptive data anonymization methods and a rapidly evolving field of research into the consequences of and potential protections against unintended data disclosures.

While this topic should influence public data disclosures, the risks and potential solutions are all contingent on the details of who would like to access the data and how they would like to use it. It has become more important than ever to understand the mechanics of potentially beneficial data disclosures, including assessment of the benefits as well as the risks. Known applications should be supported without foreclosing on the ability to develop new ones in the future.

Prior work on this topic has tended to rely on hypothetical scenarios and abstract discussions of benefits and costs. This paper takes up the challenge of characterizing potential public benefits and harms associated with smart meter data through realistic examples and finding practical ways to protect customer privacy without foreclosing on the public benefits that are associated with specific beneficial uses of meter data. This chapter argues the following:

- Consideration of potential benefits independent of potential harms or vice versa will result in sub-optimal public outcomes.
- Stakeholders, including the general public, academic researchers, businesses offering energy-related goods and services, utility customers, and the regulators and utilities themselves will have interest in the data, but the nature of those interests will be highly varied.
- The degree to which stakeholders will work in the public interest will depend on their motivations, their roles, the nature of oversight, and the opportunities for and consequences of abuse.
- Disclosure policies should include a combination of accountability for actions, legal restrictions, and technical protections on access to and disclosure of the data.
- Traditional methods of removing individually identifiable information from data still provide significant protections and should not be abandoned because they are not effective in every situation.
- Many data applications will be compatible with temporal and spatial aggregation of data and could be based entirely on pre-computed aggregate statistics.
- Data access should comply with existing privacy protections put in place by the CPUC, but such restrictions may require clarification, alteration, or extension over time in response to desired public interest outcomes.

5.3 Public benefits and harms

As a general rule, public policies related to energy infrastructure are intended to provide net public benefits. Policy changes impacting rate payers typically need to pass cost-benefit tests that show net public benefits and no concentration of harms impacting specific groups. In practice, reasonable evaluations can reach divergent conclusions, so the evaluation of costs and benefits can be a subjective process. Public utility commissions are responsible for adjudication of potential conflicts and often subject major decisions to ongoing examination and revision to ensure that benefits are maximized and harms are minimized.

This work addresses the costs and benefits of smart meter data, but it is not about revisiting the cost/benefit analysis behind the deployment of smart meter infrastructure in the first place. Rather, it is about using meter data to achieve public benefits additional to those already anticipated. This section provides an intentionally broad systematic classification and discussion of the potential public benefits and harms that might come from smart meter deployment and the use of meter data. Subsequent sections focus on the more narrow benefits and harms from the centralization, analysis, and disclosure of smart meter data.

5.3.1 Potential smart meter benefits

Increased deployment of renewable energy — Meeting greenhouse gas emission targets will require a significant increase in zero emission generation on the grid. Renewable sources like wind and solar PV are heading to cost parity with fossil fuels and are generally expected to contribute a significant portion of grid emissions reductions. Their widespread deployment would also reduce conventional air pollution, which has long been a public health problem[5]. However, the integration of wind and solar energy onto the grid will require the ability to manage the resulting natural fluctuations in available power. Depending on the amount of renewable energy deployed, operational challenges might include unexpected (or foreseeable) shortfalls (or oversupply) of generation and rates of change that exceed system ramping capacities. Smart meter data can be used to characterize system demand with finer spatial and temporal resolution than currently possible. Such data can help improve demand forecasts, identify potentially flexible users of energy, and assess the outcomes of programs designed to encourage flexibility in the timing and magnitude of energy use.

Decreased pressure for new infrastructure — The total capacity of operable power plants and the capacities of transmission and distribution infrastructure, which combined are the primary drivers of energy grid infrastructure costs, are determined by the handful of hours during the year that demand is at its highest. Among the high-profile arguments for smart meters has been their ability to facilitate the billing strategies

[5]In California, this problem is especially concentrated in the Central Valley and Los Angeles basin.

and communications necessary to roll out more ambitious demand response programs designed to curtail energy use or shift it away from the periods of highest energy use the grid experiences. If it is reliable at scale, demand response can help avoid or defer expensive capital investments.

Improved efficiency program planning — A great deal of time and effort is expended during the planning of efficiency programs to determine the magnitude and type of energy savings accessible to energy efficiency interventions. Even after decades of experience, this exercise is as much art as science, based on results from past programs and data from surveys of energy use and appliance ownership intermingled with cost-benefit economic models meant to capture consumer behavior. With smart meter data, a concrete record of consumer energy choices is now available. The patterns of usage over time and relative to various demographic and environmental factors can be analyzed to provide a more empirical assessment of drivers of consumption and waste. For example, estimates of baseloads (which notably include the rapidly expanding category of electronics known as plug loads), temperature-sensitive consumption, tightly scheduled consumption, and ad-hoc usage can be studied for geographically and temporally differentiated efficiency potential. Such information should help to plan efficiency interventions and identify the customers who would benefit most from them.

Improved efficiency program evaluation — At their conclusion and at specific points along the way, energy efficiency programs are evaluated to determine how well programs are doing at meeting their goals and to learn lessons that can be applied to subsequent programs. What type of customer participated? Did the savings persist? Did neighbors adopt similar efficiency measures? How did an intervention change measured consumption? What were the benefits realized by customers and society at large? Current best practice is to employ a patchwork of evaluation methods from surveys to econometric models, back of the envelope calculations, and intensive study of selected customers with special metering and measurement equipment. Now that smart meter data is available for nearly all utility customers, evaluation questions can be addressed using meter data in a manner that is additional to current practice or where it serves as a lower cost or more precise substitute.

Improved benchmarking — Through relative comparisons of annual energy use across many customers presumed to have similar needs, it is possible to identify customers with unusually high or low energy usage. Such information can help locate, study, and reward top performers, identify candidates for intervention, and provide context to motivate their actions. For example the EPA's Energy Star rating for commercial buildings, documented in [32], is based on energy benchmarking data from the Portfolio Manager tool. Some behavioral efficiency programs are already comparing customer usage to that of their neighbors on a monthly or daily basis. With interval meter data, many different load features in addition to annual consumption can be benchmarked. Even though smart meters record aggregate demand for an entire residence or commercial building, the patterns of that use can be used to identify more appropriate peer groups (for example customers with evidence of air conditioning or unusual hours of usage) and provide more

actionable comparisons.

Customer education — Well informed consumers tend to make better decisions than consumers with limited information. Smart meter data can provide timely access to a more detailed record of a customer's patterns of usage. Scrutiny of this data could lead customers to insights about what end uses are consuming the most energy, the recognition of unexpected changes, and better information to guide decisions about rate plans, program participation, and energy-related projects. Properly implemented access to energy data can also provide for greater transparency of billing charges, increasing the trust customers extend utilities.

Private innovation — The advent of cheap computing, data storage, and communications has inspired many entrepreneurs to dream up new products and systems of control and information management for buildings. These systems can detect system problems, anticipate or respond to occupant needs in novel ways, incorporate new sources of information in operational strategies, and help utility customers make sense of their usage or identify areas for potential savings. At the same time, existing businesses offering efficient products, efficiency-related services, or on-site generation see ways to use meter data to better identify, inform, sell to, and serve their customers. In cases like automated home energy monitoring, feedback, and visualization, products and services can be directly built around the use of meter data. In a wider variety of cases, aggregated or customer-specific data could be used to identify operational characteristics or other indicators of compatibility with new product offerings, like solar power, home automation or products that support energy efficiency and demand response.

More effective regulatory oversight — In their role of overseeing efficiency and demand response programs across the US, public utility commissions are responsible for ensuring, as of 2011 according to [34], that more than $7.5B a year of public interest ratepayer money is spent in the public interest. Regulators have similar public responsibilities related to rate design, capacity planning, transmission citing, etc. Many aspects of regulatory oversight can be improved through access to better quality consumption data. Similarly, utilities can better understand and pursue compliance goals set in terms of clear and measurable metrics developed using meter data. Under the current arrangement, utilities, the consultants they employ, and the partners they work with have far more convenient and timely access to customer data than their regulators do. This asymmetry of information would tend to disadvantage regulators in trying to understand the implications of the various potential uses of customer data and to ensure the data is used in the public interest.

Research and innovation — Researchers studying building energy use have typically had limited access to high-resolution building energy data. Constrained by limited time and budgets, many research studies have been based on annual energy totals, monthly utility bills, or proprietary existing or ad-hoc metering systems. As a result, detailed studies of energy use have tended to be case studies based on very few buildings, and studies involving large numbers of buildings have tended to document just a few energy consumption characteristics.

The information contained within smart meter energy data dramatically expands the potential scope and scale of beneficial buildings research. Studies that were formerly conducted as narrow case studies can now be conducted on a representative sample of the building stock, or even a complete sample of the building stock. Best research practices can evolve to include controlled or natural experiments and studies that were formerly restricted to average consumption or large geographic areas can become far more specific.

This means that the types of research that tend to lead to innovative breakthroughs in building operations and efficiency will get a boost from access to smart meter data. For example, there is currently a research interest in the interactions between people and the engineered systems that operate their buildings. Smart meter data can help inform studies of energy use in practice and shed light on the extent to which real world behaviors deviate from naive assumptions about occupant behaviors. Another, more immediate example is the practice of monitoring-based commissioning (MBCx). MBCx is a data-driven process for improving building efficiency developed by the building research community. Commissioning is typically done by well-trained engineers and tends to pay back quickly [76], but it has a reputation for project costs being front loaded in part because of the expense associated with the time and materials of ad-hoc metering and related information gathering. Smart meters can help commissioning engineers begin projects with more information and thus make more efficient use of their time and resources.

Lower cost and improved reliability — As a complement to the individual benefits that accrue to utility customers, the aggregate impact of such changes can be steered to produce more efficient, predictable, and flexible demand overall. That will, in turn, decrease utility operating and infrastructure costs and improve grid reliability. Such grid-wide benefits should be manifest in lower rates (or avoided increases), better quality service, and success in pursuing long-term efficiency and emissions goals.

5.3.2 Potential harms

Up-front costs — The most obvious costs associated with smart meters are the capital and labor associated with their deployment. Once meters are operational, these costs have already been incurred, but product choices and deployment strategies impact both deployment costs and installed capabilities. More relevant to post-deployment costs, meters have the potential to be modified with new features over time, so the initial investment may be supplemented by additional spending and effort in support of new capabilities.

Maintenance costs — Along with the upkeep of the meter hardware, smart meters require secure communications infrastructure from homes to their central utility, data center resources to receive, store, and process the resulting data, and security protocols and practices to protect the data from unauthorized access. They also require new processes and procedures for providing authorized access to the data. If the roll out of smart meters is incomplete, on-site meter reading requirements persist. There are also potential costs associated with foregoing the regular contact with the infrastructure in the field that

meter readers provide. To the extent that meter reading is automated, utilities can no longer expect human readers to alert utilities to situations that are potentially dangerous or otherwise worthy of investigation.

Network security — There are significant network security concerns raised by enabling so many pieces of hardware to send data into utility-owned facilities where system-critical operational controls are also housed. Conversely, there are risks associated with customers receiving communications ostensibly from their utilities but potentially generated by a malicious actor who has gained the ability to impersonate the utility. Even if the network security is not compromised, protecting the network from these risks involves substantial costs. Finally, enabling features to support programs like automated demand response involves network security risks additional to automated meter reading.

Complexity/confusion — By their nature utility services are part of the background of people's lives. The availability of reliable and cheap electricity has led customers to a blissful ignorance about the intricacies of utility infrastructure and operations. Programs like efficiency retrofits, demand response, or real-time pricing of electricity will tend to require more attention than utility customers are used to committing to their energy bills. This could lead to confusion and resentment that something previously out of sight and out of mind has become a nuisance. Even worse, people might feel that the system was designed to take advantage of them by changing the rules to work against their interests unless they are paying attention.

Loss of privacy — Analog meters read once a month provide utilities with far less information about customers' patterns of energy use (and therefore lifestyles and preferences) than interval meters. The strongly negative reaction to the deployment of smart meters by some customers likely derived from this change in their relationship with their utilities, especially for customers already feeling distrustful of the motives and practices of their utilities. It is undeniable that the installation of smart meters has created a very large store of data on utility customer energy use not previously available and that this data has real-world privacy implications. The data could be accessed by bad actors looking for specific information or patterns, mined for information relevant to corporate marketing initiatives, used by law enforcement to flag suspicious behaviors, and fed into the omnivorous National Security Administration data collection apparatus. Smart meter data can even reveal private information to other members of the same household. [71] discusses some of the network security and privacy challenges arising form the deployment of smart meters and calls for a national effort to establish customer protections.

Loss of anonymity — Related to concerns about utilities having access to additional personal information is the possibility that that information would be inadvertently or even intentionally made available to others without the knowledge or permission of customers. In the context of research and other applications of meter data, it is anticipated that legal and technical steps would be taken to anonymize, or de-identify, the data, but there are concerns about whether some techniques might be used to re-identify customer data. It is also possible that a more direct data security breach, analogous to credit card and health care data breaches that are often in the news, would leak identifiable data

directly.

Loss of control — If decisions about the accessibility of customer data are made on their behalf by utilities and regulators, there is a significant chance that the wishes of all customers, the ostensible owners of the data, will not be anticipated. In this regard, customers may feel that they have lost of control of their own data. This could be analogous to user profile data from online services or credit history impacting people in ways that they have not been able to anticipate or control.

Unwanted profiling/marketing — Most people are growing accustomed to online experiences mediated by targeted ads and content determined by their past online (and possible offline) behaviors. This can include purchase history, web browsing habits, email contents, etc. Such content is often positioned as a service to consumers, an attempt to anticipate their needs and desires, but it can also feel uncomfortably familiar and remind people they are consumers, even as they pursue social, cultural, and educational experiences. In the context of the uses of smart meter data, analyses suggestive of potential for efficiency or demand response may feel similarly invasive when presented to customers. Even worse, communications from private companies informed by meter data, for example about energy retrofits or solar installation, might provide evidence of data sharing that further erodes the trust people extend to utilities and their regulators.

5.3.3 Characterizing meter data disclosure risks

Having identified a diverse set of potential harms arising from smart meter infrastructure and data, this section focuses more narrowly on further characterizing and mitigating the concerns most closely associated with the disclosure of meter data in the pursuit of public benefits. These concerns include loss of privacy, loss of anonymity, loss of control, and unwanted profiling and marketing. This section will provide an exploration of what can be learned from meter data, how that information could be harmful, and what technical, procedural, and legal steps can be taken to minimize the risks of harm from public-interest disclosures.

5.3.3.1 Identifying relevant stakeholders

To understand the risks associated with data disclosure, it is necessary to consider the variety of stakeholders with interest in disclosures. Their differentiated motivations and capabilities define the environment for meter data applications.

Regulators are officially tasked with looking after the public interest and have a great deal of authority to pursue their goals. They set rules that shape the outcomes for all the other entities and determine where the public interest demands their intervention.

Public utilities[6] necessarily have access to complete and identifiable smart meter

[6]Here public utilities refers to state-regulated utilities that are traded on the public stock markets. The full scope and variety of utilities is being glossed over. There are, of course, municipal and other publicly owned and operated utilities as well as purely private utilities and everything in between. With

data. Their operational needs and future plans for grid management provide the foundational motivation for the deployment of smart meters. However, they are not traditional innovators in the field of information technology, and some are uncomfortable with their responsibilities and capabilities related to smart meter data. They are the public face of the grid and are most directly accountable to utility customers when problems arise. Their regulatory environment, which can be cooperative at times and adversarial at other times, often steers them toward public goals, but they are operated primarily for profit. They face financial incentives for pursuing applications of meter data outside its core uses that improve their earnings potential. Their options, including new pricing models that increase revenue or decrease pricing risks, detailed studies of customer behavior, data sharing agreements with third parties, or aggressive data-driven marketing are not guaranteed to serve the public interest. It is the responsibility of regulators to identify cases requiring intervention.

Consultants and contractors are hired by utilities and their regulators for their expertise in areas of program planning and execution, data analysis, data security, and strategic planning. They are not explicitly operated in the public interest, but they enable many public interest activities. They help set up and run efficiency programs, play important roles in short- and long- term strategic grid planning, and tend to offer flexibility and expertise that both utilities and their regulators find inefficient to maintain themselves. They have the means to perform valuable analyses of smart meter data and will often better understand that value than other entities. They also have strong incentives to use insights gained from their analyses for their own strategic benefit or to use what they have learned to retain other clients.

Private businesses with interests in meter data run the gamut from venerable energy service contractors and commissioning agents to solar companies to brand new start-ups with a dizzying array of business models related to using meter data and other data sources to offer new products and services to utilities and their customers. These companies often house valuable expertise in energy systems and data analysis, tend to be more entrepreneurial than utilities, are sources of innovation financed with private capital, and in many ways embody the potential for long-term market transformation toward clean and efficient grids. However, new companies face significant pressures for earnings and growth and have business models, informed by the success of other data-driven companies, that include the mining and marketing of personal data as a major source of revenue. Combining these pressures with operation at both technological and legal frontiers can lead them to act first and ask forgiveness later. To secure revenue and future growth, many of these companies aggressively pursue roles in publicly subsidized programs like efficiency, demand response, and grid operations and are focused on obtaining bilateral data sharing agreements with utilities.

Academic researchers are motivated by the publication of their findings and often

modest imagination, the basic motivations, roles, and lessons related to public-interest use of meter data found here can be adapted to these other situations.

explicitly work on public-interest problems. This is particularly true of public universities and publicly funded research labs. They can be expected to be significant creative forces in developing new methods of analysis and finding fault with or improvements to existing methods. Their technical expertise can match the expertise of commercial entities and therefore be of significant value to regulators and advocates of the public interest. On the other hand, some academic work can focus on theoretical objectives over practical ones and, like any creative discipline, some new academic ideas will not stand the test of time.

Bad actors are people or entities seeking personal gain at the expense of others using information gleaned from meter data. They can be considered highly motivated to find flaws in existing practices and systems so they can be exploited. Their actions could include the invasion of privacy of specific or arbitrary customers, the use of meter data to facilitate a robbery or scam, stealing utility services, industrial espionage, etc. Their actions, by definition, work against public-interest uses of meter data and will often, but not always, be illegal.

Customers are the group most intimately associated with smart meter data, but they are not often directly involved in the decisions about their data. They cannot be expected to be closely involved or well informed about all decisions, but they can be expected to have opinions as they learn what is being done on their behalf. Their frustrations related to smart meter data range from fears of disclosure of personal information, to lack of trust for the custodians of their data, to concerns that inadequate public benefits are being achieved. They are the legal owners of the smart meter data and the source of funding for public-interest utility programs. In the context of regulatory discussions about the uses of meter data, the other entities often try to demonstrate that their work is in the interest of customers.

Customer advocates represent the interests of customers in official proceedings. Unless they specialize in another area of interest, advocates tend to focus on ensuring equitable distribution of costs and keeping rates low. Their positions on the uses of smart meter data tend to focus on ensuring that benefits accrue to rate payers and that customer data is not misused.

5.3.3.2 Re-identification definition and examples

One of the most sobering risks associated with data disclosure is that clever bad actors will be able to reconstruct information that was intentionally removed from the data to protect customer privacy. In the research literature on the subject, *de-identifying data* (also known as anonymizing data) is defined as the process of removing personally identifiable information so that the underlying data can be shared without *re-identification*, which is the process of re-discovering the identity of some or all sources of the data.

For a long time it was assumed that data sets with demographic information like age, income, zip code, and gender but not individual names, addresses, or other individual identifiers could not practically be re-identified. However, a series of data releases followed by partial re-identification provided substantial evidence that de-identification is harder

than was once thought. In one well-known instance, the Massachusetts Group Insurance Commission, believing their records had been de-identified, released medical records for study by researchers and even sold a copy to industry. Using only zip code, gender, and date of birth from the medical records and public voting records containing the same data, [110] located the medical records of the governor of Massachusetts and mailed him a copy to be sure he noticed. She went on to demonstrate that the 1990 census data suggested that 87% of the US population could be identified in this manner[7].

In another example that moves beyond exact matches, [81] employed statistical techniques to compare the similarity of public movie ratings and reviews from users of the Internet Movie Database (IMDB) website with a de-identified data set with similar content provided for analysis by Netflix. These researchers were able to link Netflix activity and public IMDB profiles with very high degree of confidence for 2 out of 50 IMDB accounts studied, a 4% success rate on a very small sample. In [82], the same authors have demonstrated the ability to re-identify Twitter users using only the pattern of their connections to other users compared to a social network structure from a photo sharing site with public account data. Several other examples of similar accidental disclosures have emerged in recent years.

The pattern of these disclosures is instructive. Re-identification typically utilizes *separate public data* that contains information that overlaps with the de-identified private data. The overlap between public and private data is used to link the private data to the public data, revealing the identity of some or all members of the private data set. To be harmful in a practical sense, such methods also requires the presence of sensitive information in the private data that is not obtainable through less effort elsewhere.

5.3.3.3 Sensitive and identifiable contents of smart meter data

What constitutes sensitive data can be highly subjective, especially for meter data, which does not necessarily disclose information about individuals. The purpose of smart meter data is to capture a time series of consumption at a specific location, which is often generated by a group of people rather than an individual. Furthermore, much of the consumption is subject to the automated controls that govern heating, cooling, and appliance behaviors and may take place regardless of the presence or absence of individual occupants. The information meter data reveals is a combination of all occupants' activities and automated consumption. Still, it is possible to imagine that some industrial facilities, like data centers or specialized manufacturing facilities, could consider disclosure of even their annual total energy consumption to be unacceptable. In other cases, like efficient buildings with proud owners, public facilities, or customers seeking in advice or interested in supporting research, there may be significant customer interest in publicly disclosing detailed meter data.

Examples of residential customers likely to be sensitive about identifiable data disclosures include public figures, unusually high consumers worried that they will be judged

[7] [39] suggests that this number may be closer to 60%.

harshly, customers who want to protect their privacy, and customers using residential locations to produce energy intensive illegal goods, like marijuana, in secret[8].

From annual energy to apparent occupancy schedules, much of what might be considered sensitive about such data lies in the features that must be extracted from the time series. Many of these features are subject to irreducible uncertainties of their own. For example, algorithms designed to disaggregate end uses from hourly whole-building data have significant error rates. [54] proposed a method of disaggregation whose best configuration correctly classified about 55% of energy use starting with hourly data across nearly 600 homes and more than 10,000 appliances. Similarly, semi-physical regression models that attempt to explain patterns in meter data as a function of times of day, days of the week, observed variations in weather, and other data on potentially causal processes explain a modest fraction of energy demand variations but also have significant uncertainties in their estimates. Such algorithms can be expected to perform better better when applied to data captured at finer time scales, as the distinctive patterns associated with individual loads are revealed.

Assuming a customer has been identified and a bad actor has a workable model to link features to sensitive information, the most accessible meter data features include:

- Annual, monthly, and daily totals of energy suggestive of occupancy or operation of large appliances.
- Daily load shape characteristics suggestive of patterns of equipment use and occupancy.
- Weekly and seasonal patterns of energy use suggesting work schedules or long absences.
- Estimated patterns of occupancy, both in general and for specific time periods.
- Timing of unusual events that might coincide with prior knowledge of context sensitive activity.
- Estimates of end uses or the presence or absence of specific end uses[9].

Cases where the energy data features are themselves the secrets people want to keep, as would be the case for industrial espionage or embarrassing levels of consumption, should be distinguished from cases where the patterns of energy consumption are indicative of other secrets, as would be the case for the employee pretending to be home sick while

[8]Naturally, law enforcement agencies can and do request access to the energy data in cases of suspected criminal activity, but more public disclosure of data could bring about suspicion in the first place.

[9]It is not possible to anticipate all the special cases of features of meter data that might be of use to a hypothetical bad actor under hypothetical circumstances. An example provided in a Microsoft Research white paper postulates a utility customer whose religion is a high-stakes secret that could be revealed by patterns of energy use around specific religious holidays.

instead secretly attending a baseball game. It is also worth noting that these features become more difficult or impossible to reconstruct as data is aggregated in time or across multiple customers and that some information, like current occupancy, that is considered unacceptably revealing can become less sensitive with the passage of time.

In addition to time series observations, utilities have account information, including account number, service and billing addresses, credit card or bank information, service hookup type (i.e., line capacity and single vs. multi-family hookups), date of meter install, rate plan, etc. Obviously this data uniquely identifies customers, but the current practice tends toward revealing zip code, service type, rate plan, and dummy account IDs that preserve cases where multiple locations are billed to the same account. Assuming best practices for data disclosure have been followed and the meter data alone is being used to attempt to re-identify a customer, this identification will very likely rely on the public disclosure of features similar to the above list. Examples of public disclosures that might enable a link to anonymous meter data include:

- Prior disclosure of identifiable annual or monthly consumption, which could be matched against annual or monthly totals from meter data. Some cities require such disclosures from their larger commercial properties.

- Prior disclosure of some subset of data that can be directly matched to the whole set. Particularly motivated bad actors might gather this data personally on site, or some customers may reveal data samples for educational or illustrative purposes.

- Public knowledge of unusual events, like outages, large gatherings, or specific dates of changes in occupancy that could be correlated with related patterns in meter data.

- Public knowledge of extreme consumption (high or low) that could identify the only possible customers capable of consuming at these levels.

- Public knowledge of unusual operational details. For example, a refrigerated warehouse or ice rink will have predictably high fractions of energy correlated with outside temperatures, and a building operated with an unusual schedule could similarly be identifiable. In the residential context, it is possible that pool, hot tub, or solar panel ownership would be sufficient on their own or in conjunction with other characteristics to reveal identity within certain communities. Even the presence or absence of central air conditioning, heat pumps, electric heat or hot water could be unusual in the context of an identifiable group. Similarly, a night shift worker might be exposed by hourly meter data out of step with typical diurnal patterns.

- If disclosed, service and rate types, particularly those that are rare, could be used to narrow in on the identity of a customer. For example, rates associated with solar panels, low income, and specific residence types might stand out, especially in combination with other known features.

- In cases of multiple locations being tied to a single account, or master meter accounts for multi-family residences, it may be possible to to deduce which locations fall into such unusual categories.

At present, it is not known how often these theoretical privacy threats will surface as real-world problems. Much of the information that can be learned approximately from meter data can also be learned exactly through similar effort applied elsewhere. It is also not known what portion of the population has active concerns about the privacy of their meter data. We do know that push-back against smart meter installation has come from a minority of customers and has been stronger and more persistent than anticipated.

If experience with other forms of consumer data is a guide, the incorporation of energy data into the growing body of information used to study, track, and sell to consumers is likely to be the *most widespread and consistent* privacy concern raised by meter data. Unlike the acts of criminal surveillance discussed as bad actor risks, the commercial use of meter data is not necessarily illegal or even undesirable. Many commercial uses, for example identifying households to approach with solar installation or energy retrofit offers, have significant overlap with utility program implementation and larger market transformation policy objectives. However, meter installations are explicitly performed for the public benefit, and installation is virtually mandatory. Under these circumstances, the concerns of customers reluctant to share their data or see it used for commercial purposes should carry substantial weight.

The default protections extended to passive customers should protect them from both bad actors and unwanted commercial profiling. Regulators forbid utilities from directly selling customer data, but there is a large gray area that allows customer data to be shared with a significant number of companies in the course of utility operations or program design, execution, and evaluation. At least in California, this commercial data-sharing exception is currently the primary mechanism of utility data transfer. Put another way by [69], *"because many state regulators have an interest in leveraging spending by program administrators and/or are interested in transforming markets over the long term, this issue of access to energy efficiency-related customer data to support private sector business models is an important policy issue."*

5.3.4 Current state of practice

This section describes current customer data sharing practices in California in the context of the characteristics and disclosure risks detailed above.

There is no question that utilities have operational and planning processes whose outcomes can be improved through the use of meter data[10]. These applications are among the most natural uses of the data and are routinely permitted by regulators. Outside these core uses of meter data, standards for data access, including public interest research, are

[10]In some jurisdictions, responsibility for efficiency programs or other utility programs are delegated to third parties that are also granted access to meter data as a part of their duties.

variable from one jurisdiction to the next but are typically less permissive. As a result, most meter data access is currently negotiated through bilateral data sharing agreements between utilities and third parties overseen by regulators.

5.3.4.1 Unintended consequences of current practice

The general pattern of utilities brokering permissions and access to meter data is an understandable delegation of responsibility by regulators. For example, California Public Utilities Code Section 8380(e)(2) authorizes utilities to share customer data with third parties *"for system, grid, or operational needs, or the implementation of demand response, energy management, or energy efficiency programs"* without customer consent or knowledge. The effect of this policy is that the entities with the most permissive access to meter data are utilities and the consultants that serve them.

According to [103, 101, 89], California's three public electric utilities jointly reported sharing identifiable customer data with a total of 367 authorized third parties in 2012. Third parties can include contractors, consultants, municipalities, businesses that directly engage utility customers, and researchers. However, the reported numbers were not broken down by category [11].

No violations of contractual provisions were reported, but in the course of their business activities, the group of companies that routinely handle customer data can face strong temptations to engage in the activities that motivate protections of customer data in the first place. Many uses, like the development of internal market research and strategies based on meter data, would be next to impossible to detect or regulate from outside these companies.

The risks of off-label and ambiguous uses of the data and of accidental disclosure increase with the number of third parties with access to the data. Indeed, some of the custodians of such data, including companies and their individual employees, have significant business interests in private information and strategic insights gained through the analysis of meter data.

Particularly at companies that have a specialized focus on analyzing customer data, potential uses of the data abound. For example, energy data feedback pioneer, Opower, maintains a blog, called Outlier, based on its explorations of identifiable customer data from 50 million utility customers tied to external data on location, household, and demographic characteristics. In the past it has used the data to examine methods for predicting human behavior, who left home for Thanksgiving, and how consumption is influenced by home size, political affiliation, email client choice, and public voting records[12].

The relatively open access to data enjoyed by utilities and their consultants stands in

[11] At least 57 of these were vendors directly authorized by customers to access their data. Some are likely to be municipalities or other public entities with interests in local energy use. Many were surely related to the planning, implementation, and evaluation of public-interest utility programs or grid planning activities. but the actual mix was not made public in the official filings.

[12] All links last accessed 9/20/2013.

sharp contrast to tight restrictions on data for researchers, program implementers, municipalities, and solar and efficiency contractors. California Public Utilities Code Section 8380(b)(2) prohibits the sale or transfer of *"customer-specific energy usage data to third parties for commercial gain or profit-making purposes"* including solar promotion, energy efficiency retrofits, loan product marketing, and on-bill financing. The sale or transfer of aggregated data is only permitted on a case-by-case basis with approval from the utilities commission. According to [69], this tiered access can hamper the progress of efficiency programs. They observe that *"the inability of energy efficiency service providers (EESPs) to gain access to the data because of legitimate privacy concerns creates a barrier to realizing many of the benefits from these services."*

The reason for this asymmetric arrangement is that utilities need to share data with their delegates to conduct their business, and the other activities are peripheral to utility operations. From a privacy perspective, however, the effect is that inconsistent standards are being applied to customer data. Identifiable data is routinely shared with hundreds of *authorized third parties* through utilities each year but has remained out of reach for researchers and companies that install solar panels or perform efficiency retrofits due to privacy concerns. Where privacy is concerned, the weakest protections determine the effective strength of the protections. Care should be taken to apply methods for protecting customer data to all privacy risks.

5.4 Protecting customer privacy

This section focuses on methods for protecting the private information in utility customer data. The standard tools of privacy protection include legal agreements, data minimization (providing only the data needed for a particular task), anonymization (removing personally identifiable information), and data aggregation. In response to learning about cases of re-identification in other fields, the report from the Working Group for Phase III of Rulemaking 08-12-009 in California, [115], stated that *"expert technical solutions are required to develop more robust privacy solutions because current anonymization and aggregation techniques may fail to protect private customer data."* They identified *"finding an acceptable, reasonable 'solution' to the 're-identification problem"'* as a *"condition precedent to development of a practical, standardized process for energy usage data access in this proceeding."*

Unfortunately, one of the main lessons of the re-identification literature is that there are no universal protections against re-identification. The ability to re-identify an individual is contingent on an external source of information whose characteristics determine the feasibility of and path to re-identification. This section explains why statistical tools that add noise to data, like differential privacy, cannot provide a practical solution to this problem. It offers a pragmatic suggestion to differentiated access to meter data according to motivation and ability to skirt privacy rules and advocates implementation of a system of *delegated analysis* that would allow stakeholders to specify the details of the analysis

they would like to perform and gain access to the results without requiring access to the source data.

5.4.1 Differential privacy

Differential privacy provides techniques for adding noise to data in a manner that preserves some features of the underlying data while obscuring others. It was developed in response to the challenges posed by re-identification and has enjoyed substantial attention in California's deliberations on data privacy. However, it is no panacea. In this section, we provide a discussion of the general concepts, strengths, and weaknesses of a differential privacy approach to the challenge of protecting the privacy of meter data.

The theme of independent public data enabling the re-identification of private data has not been lost on privacy researchers. Researchers have proposed many creative systems designed to protect customer identity or avoid data disclosures in the first place. For example, [27] defines differential privacy and derives a method, described below, for ensuring it. [98] advocate using cryptographic techniques to perform data feature extraction on-site, ensuring limited off-site data disclosure and tamper resistance of the underlying calculation. [29] describe techniques for pooling meter data via an escrow service to mask identifying information while providing high-resolution and timely data in support of grid management. [106, 58] describe cryptographic and differential privacy techniques that allow reconstruction of time-series totals across several sources without the ability to determine individual contributions. Of these systems, differential privacy is the most relevant to generalized data protection and has specific and computable privacy guarantees, so it merits further discussion.

The key breakthrough that enabled the development of differential privacy was a new definition of privacy. One potential definition of privacy is that anything that can be learned from data in a public database could be learned without the database through other means. This sounds like a good goal in theory, however, a database with useful or unique information cannot provide this guarantee. [27] re-frames the problem in terms of the risk of re-identification run by an individual deciding to allow his/her data to be added to (or removed from) a database. For that individual, *differential privacy* is achieved when two separate samples of the data differing by only his or her entry do not have statistically significant differences in their properties.

Differential privacy can be shown to support rigorous computable privacy guarantees. Specifically it introduces noise into the underlying data in a manner that guarantees that the ratio of the probability of observing an original sample of data to the probability of observing the same sample with one row added or removed is no greater than a predetermined threshold. Specifically, with data rows consisting of a single presence/absence piece of information, noise drawn from a Laplace distribution (which is two exponential decay functions back to back) with a scale parameter $1/\varepsilon$, written as $Lap(1/\varepsilon)$, will keep the probability ratio below e^{ε}, where ε has been determined in advance and can be known to the public.

If N separate queries of the data will be made, or if N pieces of data are different between data sets, differential privacy can be recast as a privacy budget spent across all the data accessed, with noise of the form $Lap(N/\varepsilon)$ added to the data returned for each query result. In other words, the scale of the noise increases with the magnitude of potential difference between data samples to cover those differences. The Laplace distribution flattens out and grows wider as N/ε, increases. The flatter the distribution, the greater the probability of adding large disturbances to the underlying data, potentially rendering it useless for productive analysis. Figure 5.1 on page 156 illustrates this effect for $N = 1$ and various reasonable values of ε. The area under each curve is fixed at 1, and the area under the curve between two values on the x-axis is the probability of a value in that range being drawn from the distribution.

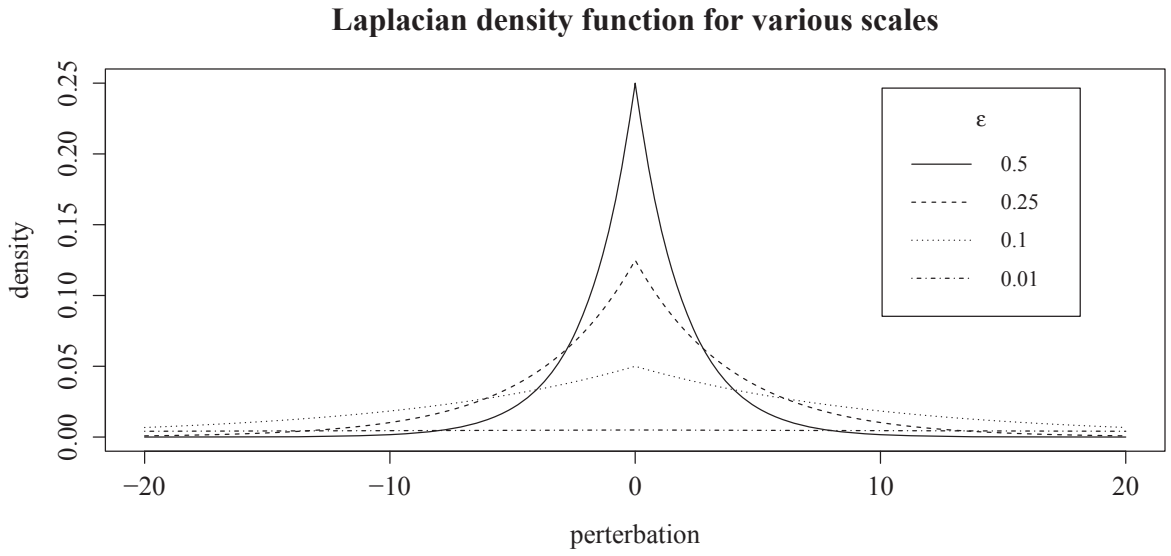


Figure 5.1: Laplacian distribution for $N = 1$ and various values of ε in the range suggested by [26].

Utility meter data tends to have many time-series observations per customer (i.e. large N), so meter readings will require very large disturbances to enforce data privacy. The tradeoff between the ability to do useful analysis, particularly analysis that is open ended or exploratory in nature, and the constraints of differential privacy cannot always be resolved. [26] is pessimistic about applying differential privacy to data intended to support open-ended exploration: referring to a *"method of generating a 'noisy table' that will permit highly accurate answers to be derived for computations that are not specified at the outset,"* she says, *"the noise bounds say this is impossible: No such table can safely provide very accurate answers to too many weighted subset sum questions."* Going further, [14] offers a proof that a universally optimal tradeoff between utility and privacy cannot

be found for practical cases. In other words, there is no such thing as general-purpose differential privacy. It requires that you know how you will use the data in advance so it can be implemented to preserve the features of the data required to support that use.

Even more troubling for meter data, it is unclear how well differential privacy can work for time-series data. The sequence and relative position of time-series meter readings have special meaning because they capture processes with continuity over time. Noise added independently to each reading can erode the underlying time continuity of readings. Work on the application of differential privacy to time-series data has made some progress. In one example, [95] demonstrates a method that applies Fourier transforms to N time series observations to reduce the number to $K < N$ Fourrier coefficients, but useful applications of meter data could easily require K to be impractically large as well. In the special case of constructing a time series that represents the sum (or product) of observation from many households, [106] provides a differentially private algorithm but points out that the novel approach leaves many questions unanswered.

At least for now, differential privacy is not a practical model to protect data with many observations per person, observations organized in time series, or data intended for exploratory research. It is best suited to data containing a modest number of discrete values, like true/false or low integer values. Examples might include an indication of whether a home uses more or less energy than average or what billing tier a home falls into. Because many of the potential benefits of utility data rely on more complex characteristics, it appears that differential privacy cannot solve the central problem of ensuring the privacy of utility customers while preserving the usefulness of the underlying data. It could still prove useful in cases where only basic summary statistics will be made available through filtered interactive queries.

Caution is also merited because the field of research into differential privacy is just taking shape. Tested and reliable commercial implementations of differential privacy compatible with meter data are simply not available. Future work could overturn some part of the foundation upon which the guarantees of differential privacy rest. Even if the theory holds, a specific implementation of differential privacy could be found to have fundamental flaws. This is a concern that often arises in software that utilizes random values. Most computers are simply not capable of generating truly random numbers.

5.4.2 Practical privacy for meter data

The fact that differential privacy is a poor fit for many applications of meter data does not mean that the examples of re-identification that motivated the development of such protections should not be taken seriously. There are other viable ways to protect the identity and sensitive information of utility customers in practice. These include data minimization, anonymization, legal agreements, tiered access to data, data aggregation, and extending privacy controls to utility customers. In fact, most of the practices that existed prior to the advent of re-identification techniques are effective deterrents to re-identification techniques. For example, Ann Cavoukian, the Information & Privacy

Commissioner of Ontario, has assembled data privacy best practices into a concept she calls Privacy by Design. She has written many reports on the subject, including, [18], which presents her work for San Diego Gas and Electric on applying the principals of Privacy by Design to third party access of customer data. This work is not based on Privacy by Design, but shares some of its common sense prescriptions.

5.4.2.1 Data minimization

Recall that re-identification relies on a public source of information with information that corresponds to information in de-identified data. If a bad actor will rely on census data or voting records for his attack, he will need information found in those data sources to be in the meter data as well. If good data anonymization and minimization practices are followed, extraneous information will not be present in the data to support re-identification. For example, in combination, gender, birthday, and zip code might reveal the identity of many individuals; zip code alone will not.

One of the surest ways to ensure that information about customers is not revealed is to not collect or store it in the first place. [72] suggests that *"personal data should be collected only for specified purposes, that data should be adequate and not excessive, and it should be kept in a form allowing identification for no longer than strictly necessary."* In the context of meter data, there is little question that utilities will require addresses, billing information, and other personal and identifiable details. However, data supporting the work of authorized third parties and other interested parties could be drawn from a pre-minimized and pre-anonymized version of the data. This practice would reduce the number of identifiable copies of customer data in circulation and reduce the temptation and feasibility of intentional misuse of the data. Identifiable data could still be provided to researchers and third parties whose services require access to account-specific information, but these instances would be fewer in number and easier to monitor than the distribution of identifiable information under current data sharing practices.

5.4.2.2 Data aggregation

[69] describes a 15/15 rule used by some utility commissions to protect data privacy. A 15/15 rules requires that data be summed across customers such that no disclosure draws upon fewer than 15 customers and no single customer contributes more than 15% of the total energy. Some utility commissions take this further and only allow disclosures at the level of whole towns or only for monthly time intervals. These prescriptive protections lack the formal privacy guarantees achieved by differential privacy, and privacy researchers worry especially about data aggregations whose membership is otherwise consistent over a time period during which one member is added or removed. This might happen when residents move into or out of a home. In these instances, that member's consumption will be revealed by the difference between the data with and without their contribution.

As a practical matter, properly implemented data aggregation is likely to protect

customer identities in most cases. The operative question is what information is lost in the process. Many beneficial applications of meter data will be incompatible with aggregation or would benefit from a different metric of aggregation. For example, solar or efficiency providers might be interested in rough counts of the number of customers in the highest tiers of electricity demand or the mean cost of utility bills or some metric of variability in consumption across customers. None of these data requests would reveal any more customer data than average demand does. However, it would be difficult to anticipate the range of potentially useful metrics that might be requested.

5.4.2.3 Legal agreements

Enforceable legal contracts prohibiting the misuse of customer data are, of course, still relevant to protecting customer privacy. Re-identification changes the definition of what might be legally considered personally identifiable information, but the legal tools used to protect identifiable data have long been in place. Some scholars suggest that more emphasis should be placed on understanding the individual motivation, capabilities, and trustworthiness of data recipients. [86] frames the revelation of re-identification techniques in legal terms, providing a thoughtful discussion of the legal implications of re-identification techniques undermining the reliability of traditional protections of anonymity. He focuses on the irreducible tradeoff between utility of data and privacy protections and suggests relying on more private forms of disclosure, establishing trust relationships, and more carefully evaluating potential motives for re-identification or other abuses of private data before sharing data. [83], as leading practitioners and scholars of re-identification, affirm that *"privacy protection has to be built and reasoned about on a case-by-case basis"* and technological protections must be accompanied by *"non-technological protection methods such as informed consent and contracts specifying acceptable uses of data."*

5.4.2.4 Customer rights and controls

Examples of potential privacy violations make clear that what constitutes sensitive information is fundamentally subjective. Only customers can truly judge what revelations are acceptable. Regulators and utilities are custodians of customer data, entrusted to use that data in the public interest. They should therefore work toward giving customers greater access to their data and improved transparency of and control over how their data is being used.

Through the open data transfer protocols Green Button and Green Button Connect, steps are already being taken to provide customers access to their data and control over granting third parties access to it, but more can be done. As the owners of their meter data, customers should be able to prevent it from being used for any purpose they object to outside mandatory applications to billing and other account-level tasks. As is required for online marketers, opt-in and opt-out controls should be given to utility customers. Such controls should differentiate between data used for utility operations,

regulatory enforcement, public-interest research, public data summaries, and commercial analysis. The default options should reflect an understanding that many customers will passively accept whatever the default values are. Customers should also have access to information on who is accessing their data, for what purposes, and an accessible interface for controlling access and requesting deletion of data not required for utility operations. Agreements for sharing data should include clear provisions on the circumstances under which data will be deleted and specify processes for propagating and executing customer requests to opt out of analysis or delete data.

In the spirit of recognizing and formalizing the rights of customers over their data, some advocates have begun to articulate concepts for a customer bill of rights. For example, Limor Fried, a leader in the development and use of open, affordable, and accessible electronics has seen many useful applications of networked, communicating electronic systems in the so-called Internet of Things. She has recently proposed the development of a consumer bill of rights for users of networked data gathering systems[13]. By analogy, a clear articulation of utility customer rights should help guide future privacy practices and regain the trust of customers skeptical of the motivations behind gathering their meter data.

5.4.2.5 Differentiated access

As described in section 5.4.2.3, some researchers have suggested that data sharing agreements should be shaped by an understanding of the individual motivation, capabilities, and trustworthiness of data recipients. With no other information to indicate otherwise, it is reasonable to assume that bad actors willing to exploit specific information about individual utility customers will be found evenly distributed throughout the population and that accidental disclosures could happen through anyone with access to the data. These risks should be managed through data minimization, training, oversight, accountability for outcomes, and tiered access to data that ensures relatively few trusted people and organizations have access to sensitive data.

However, there are also systematic incentives and motivations for the use and misuse of customer data associated with each type of stakeholder. California's prohibition on the sale of customer data by utilities and trusted third parties and the restrictions on service provider access to data reflect an awareness that there is money to be made using customer data for commercial purposes that are not in the public interest. Section 5.3.3.1 details several broad categories of stakeholders with differentiated roles and relationships with the use of data in the public interest. Section 5.3.4.1 highlights the privacy risks posed by third parties with legitimate access to the data.

Differentiated access, which means providing access to different levels of data detail to stakeholders based on their needs, capabilities, and motivation to work in the public

[13]See for example her Room for Debate opinion piece from 9/8/2013 in the New York Times: http://www.nytimes.com/roomfordebate/2013/09/08/privacy-and-the-internet-of-things/a-bill-of-rights-for-the-internet-of-things.

interest, could address stakeholder-specific potential for misuse of data. As a practical matter, the rules of differentiated access could be too tight to achieve legitimate and beneficial uses of data or too loose to prevent misuse. This is where the accounting of public benefits and harms comes in. Some forms of data exploration, like detailed analysis of identifiable data could be judged sufficiently beneficial to allow the data to be gathered, stored, and shared, despite the risks of disclosure. This should be particularly true when the parties involved are trusted to work in the public interest and are taking all reasonable precautions against disclosure. For example, researchers who will help to discover new beneficial applications and characterize and improve existing practices might be given access to identifiable data under some circumstances.

5.4.2.6 Delegated analysis

One of the biggest drawbacks to differentiating access to customer data is that stakeholders with serious potential and intention to serve the public interest can be denied access to the data required to carry out their intentions. Ambitious climate change mitigation, energy efficiency, renewable energy, and demand response goals will all require innovative and creative solutions to currently unresolved problems. A system that too severely or rigidly restricts access to demand data could stand in the way of these outcomes.

Differentiated access to data can be thought of as setting an upper limit on the level of detail available to stakeholders. Because most forms of data analysis revolve around one or more specific questions or metrics, the data inputs to an analysis tend to be more detailed than the outputs. This means the data inputs rather than the outputs will run up against access restrictions[14].

To maximize the flexibility of privacy respecting analyses, we propose a system of *delegated analysis* that allows a trusted party, i.e., the Public Utility Commission or their delegate (which could be an automated system in some cases), to access and perform analysis on protected data on behalf of a less-trusted party. This approach isolates privacy risks to the outputs of the algorithm, rather than the inputs. Provided the outputs are processed into a non-identifiable form, the less-trusted party could gain the information it seeks without ever accessing data that reveals more information than it is trusted with or requires. As long as the trusted party is able to verify that it does not lead to unacceptable data disclosure, the less-trusted party could even specify the algorithm used to perform the analysis. Such a system, diagrammed in Figure 5.2 on page 162, could also be used to produce standardized public reports based on customer data on a

[14]By definition, any output that can be derived from specific inputs will be no more revealing than the inputs themselves, but as a practical matter it could be argued that a technically challenging estimation of sensitive information could make that information more accessible than it was in the raw data. It is also possible to imagine an entity with proprietary access to additional household information whose analysis outputs would expose more information that available through the meter data alone. The point here is that the inputs and outputs can be characterized separately, and the outputs, if they are designed properly, could disclose less and be shared more widely than the inputs.

regular basis. The work of [72] corroborates the viability of this approach under more narrowly defined circumstances. They conclude that personal data can be minimized for most applications involving *"system balancing, demand reduction, demand response and distribution network operation and planning."*

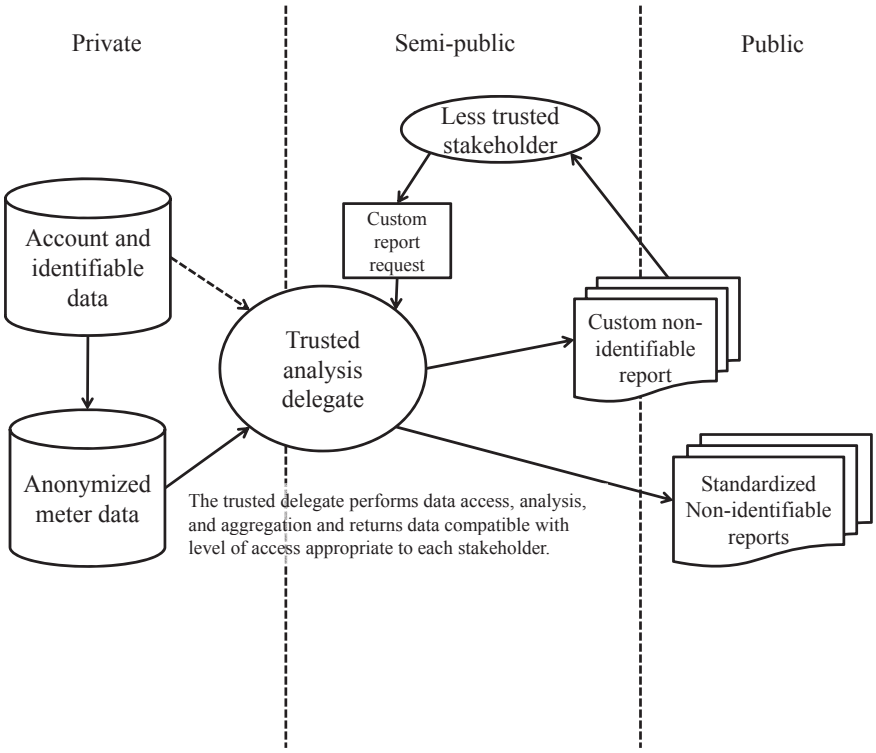


Figure 5.2: Diagram of the process of delegated analysis.

As an example of delegated analysis in areas with increasing block rates for electricity, consider a solar installer who wants to know which areas have concentrations of customers who routinely reach the upper tier with the highest per kWh rates. These are the people for whom the economics of solar installation are most likely to pencil out. The standard aggregation technique of summing monthly or daily data within a zip code (for example) will tell the installer very little about where its services are needed. If the installer can instead request a run of an algorithm that totals the number of kWh each customer consumes each month, overlays the block rate pricing, and returns zip code aggregated percentages of homes reaching the top tiers, it will have valuable, but not personally identifiable information that helps it learn where to look for new customers.

If necessary, the trusted delegate can ensure that the data returned meets any verifiable standard of anonymity. For example the 15/15 rule could be enforced, zip codes with too

few customers could be dropped, or noise generated to ensure differential privacy could be added to the returned data.

As an added benefit to regulators and utility customers, a system of delegated analysis would centralize data requests and analysis. This would allow for an accounting of how customer data is being analyzed and by whom. Regulators could choose to release the outputs of analysis and/or the details of the request, including the algorithm used, to the public. Such actions would enforce public accountability, broaden the public benefits of the analysis performed, propagate innovative analyses of data, and discourage abuse of the system.

5.5 Examples of delegated analysis

This section provides more examples of potentially beneficial analyses that can be performed using delegated analysis. We begin with figures that visualize some of the information that can be derived from smart meter data and conclude with a table of example analyses grouped into categories relevant to specific stakeholders.

Figure 5.3 on page 164 presents a curve of cumulative annual demand for electricity from a representative sample of residential customers in the service territory of Pacific Gas and Electric (PG&E), which serves most of northern and central California. The darker cumulative curve is computed by sorting the annual usage from every home from highest to lowest before summing the values. The dashed lighter cumulative curve is calculated by summing the same data in random order. The difference between the two curves can be used to calculate the potential improvement in utility program effectiveness that could be achieved through targeting. For example, a hypothetical efficiency measure that saves a fixed percentage of total annual energy and is performed in 20% of homes would produce twice the energy savings if those homes are selected from the highest annual consumers rather than selected at random. In other words, the value of the sorted line is twice the value of the dashed line at the 20th percentile home. These curves, or statistics derived from them, can be computed by a trusted delegate and returned to planners or service providers.

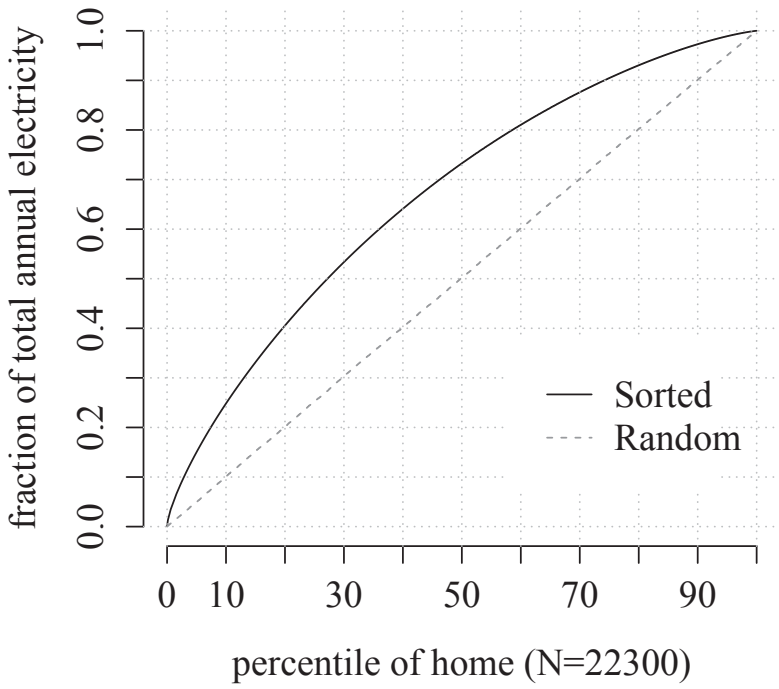


Figure 5.3: Cumulative distribution of residential annual electric energy, sorted from highest to lowest residence compared to cumulative density of annual energy based on random selection from the population.

Figure 5.4 on page 165 presents two maps that display calculated values color-coded by zip code. The map on the left shows the zip code average demand coincident with the top 30 hours of grid demand. In other words, it is a map of who contributes the most to peak demand in PG&E's service territory. System peaks in California are driven by cooling loads during the hottest days of the year. The zip codes with the highest demand are in the Central Valley, where summer weather is hottest. This result is consistent with expectations but also provides detailed and actionable information about where specifically the highest demands can be found. Those areas have the greatest potential for shaving cooling demand during system peaks, thereby reducing the pressure on building new capacity and other expensive infrastructure.

The map on the right presents zip code averaged values for estimated annual cooling energy. Estimates of cooling response to temperatures were made using a regression model that fits daily electric energy consumption with a piecewise linear temperature response. The cooling response coefficients were used to get model estimates of cooling loads for

every day of the year, which were then summed to produce estimates of annual cooling energy. This map also confirms significant cooling demand in the Central Valley, with results relevant to planning and targeting HVAC tune-up or replacement programs.

Peak coincident residential demand, averaged over 30 peak hours

Estimated annual residential cooling energy

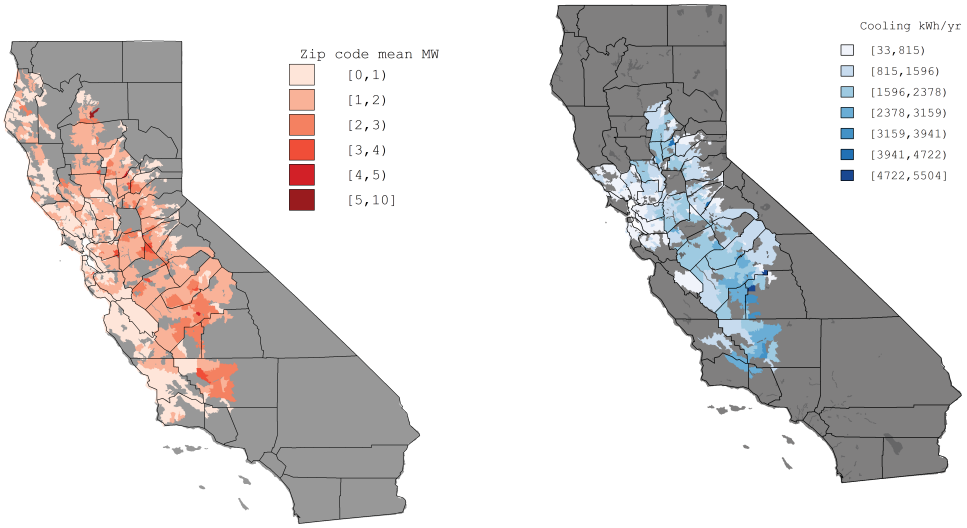


Figure 5.4: Maps of grid peak coincident demand, averaged by zip code over 30 peak hours across 3 years (left) and zip code averaged estimates of annual cooling energy, based on regression analysis (right).

Table 5.1 on page 166 provides many additional examples of potentially beneficial analyses involving meter data that could be performed using delegated analysis. The analyses are separated into various categories, each relevant to a different type of stakeholder. They are intended to provide a sense of the breadth of potential benefits achievable using smart meter data.

Table 5.1: Examples of analyses that could be performed using delegated analysis for different categories of usage.

Category	Purpose	Algorithm description
Program planning		
Inferred building stock characteristics	Recover building stock information similar to what is found in building stock surveys like [31], [30], [47], and [51]	Several basic metrics, related to the magnitude and variability of usage, aggregated geographically and produced seasonally or annually. Techniques that estimate end use breakdowns or the presence or absence of specific loads, like heating and cooling and large appliances, could also be used.
Spatial distribution of demand	Provide information about spatial patterns in demand for program targeting	Zip code averages of total monthly, seasonal, or annual energy use. Alternately, similar spatial treatment other basic demand metrics, like daily minimum or maximum demand, variance in demand, or estimated seasonal heating or cooling energy.
Energy Efficiency and Demand Response potential assessment and targeting		
Inferred individual building characteristics	Estimate heating and cooling loads, base loads, and variable loads.	Regression models run on each customer's time series meter data, with external weather data matched by zip code. Regression coefficients and performance metrics corresponding to heating and cooling loads, base loads, and load variability averaged by zip code.
Utility expenditures	Estimate bill totals and tiered usage breakdowns	Use hourly meter data and rate-plan information to calculate approximate bill totals and breakdowns by billing tier or other billing categories. Report averages of these values for zip codes to support assessment of savings potential. This assumes actual utility billing data would be out of reach for the trusted delegate.

Demand at peak	Provide cumulative demand curve during system peaks and rough locations of highest demand	Date and time of peak demand periods are used to locate customer demand data coincident with grid peak. Customer demand used to make a cumulative distribution of total demand, sorted from highest to lowest, and zip code averaged demand values.
Municipal use		
Basic consumption reporting	Provide regular consumption statistics for the municipality	Regular updates to aggregated basic consumption statistics (means or sums) across a full municipality, or across districts within a municipality, most likely defined via lists of zip codes.
Measurement of ordinance impact	Measurement of changes in bulk demand characteristics before and after the implementation of an energy-focused ordinance	Calculate baseline energy usage, aggregated at the municipal level and controlling for weather trends and other potentially confounding factors, before implementation of an energy ordinance, and calculate the same values after implementation. Return the difference between the two.
Informing the marketplace		
Standardized reporting	Public disclosure of industry-relevant summary information on a regular basis	Periodic runs of fixed algorithms, designed to provide industry-relevant information over time. Algorithms would be determined by consensus or selected from the others found here. Likely candidates would be monthly, seasonal, and annual demand, along with estimates of end use categories, averaged by service territory or zip code.
Correlations between other characteristics and utility expenditures	Support for prediction of energy use based on census data	Calculate covariance or regression coefficients between zip code aggregated census data and zip code mean customer consumption metrics.

Locating solar benefits	Help solar contractors find areas with customers likely to benefit financially from installing solar panels	Using customer meter readings and their rate plan, calculate the portion of their expenditures under each rate type. Return the average of these values or the percentage of homes beyond a specified threshold by zip code.
Program evaluation		
Bulk energy/aggregate changes	Measurement of changes in bulk demand characteristics before and after a program	Calculate total energy demand for fixed time periods and fixed geographic areas or lists of customers before and after a widespread program intervention to determine impacts. Actual calculations may require weather, trend, and other corrections to individual customers before they are aggregated.
Controlled experiments	Track changes in energy across a control and treatment group	Define treatment and control groups of customers and estimate bulk demand changes for each group over the same period of time. Calculate the difference of the bulk changes to find the treatment effect.
Grid operations and management		
Improved bottom-up forecasting models	Train forecast models using customer data	Use customer data aggregated to the substation, distribution line, or other grid-relevant point of aggregation to train a forecast model of customer demand. Report spatially distributed forecast model parameters.
Geo-targeted efficiency and demand response	Correlation of demand pockets with locations of grid congestion	Use operational data to locate areas of grid congestion or otherwise stressed infrastructure. Query meter data for customers served by (or nearby if actual service connection information is not available) stressed infrastructure and analyze data belonging to these customers for high consumption, returning zip code level counts of homes well suited to high-priority energy efficiency and demand response interventions.
Customer education and empowerment		

Billing analysis and advice	Provide customers visibility into their options for reducing bills	Using customers' access to their own data, calculate various factors contributing to demand charges for past consumption and run usage reduction scenarios to provide estimates of the impacts of efficiency, conservation, demand cresponse, and solar.
Benchmarking against peers	Provide customers a sense of how they are performing relative to their peers	Using customers' access to their own data, and a search criteria that returns a population of similar customers, compare total energy and other usage metrics, like baseload and variability, to population-level benchmark distributions.
See the impact of efficiency measures	Provide customers access to a before-and-after usage comparison	Using customers' access to their own data, build a prediction model for demand, controlling for weather effects and trends, using data from before efficiency measure implementation and possible correlations with other customers. Subtract model predictions from measured outcomes to calculate energy impacts of efficiency measures.
Frontiers of research		
The development of new analytical algorithms	Research into new algorithms that could be used as a part of delegated analysis	The execution environment used by the trusted delegate to perform analyses for others could also provide an environment for researchers to develop and test new algorithms. Such algorithms might include measuring the efficacy of energy efficiency and demand response measures using the expected timing or other characteristics of savings, testing and improving methods for using meter data to support program evaluations, better understanding and characterizing the contribution of occupant behaviors to loads, developing and improving methods of disaggregating end uses from total demand, developing methods to diagnose efficiency potential based of patterns of consumption, etc.

5.6 Discussion

There is an inherent tension between making the most of smart meter data and protecting the privacy of utility customers. When discussing the highest use of meter data, accounting for both the potential benefits and harms, it is important to judge proposed changes relative to the current practice as well as against absolute standards. Compared to current practice, there is significant room for making more productive use of meter data analysis while reducing risks to customer privacy. This will not require achieving a theoretically perfect system. In fact, it is unclear that the system could be perfected in this context. The very existence of meter data introduces privacy risks to customers. However, there are many tools available to manage those risks.

The relatively recent advent of sophisticated re-identification techniques that use correlations with external data, like voting records or online activity, to make educated guesses about the identity of the person responsible for de-identified data has created significant uncertainty over how much privacy protection is enough. The bad news is that we can never be certain there isn't some external data out there that is perfectly suited to the task of re-identifying protected data. The good news is that all of the tools and techniques previously available to protect data are also effective against re-identification. Data minimization, anonymization, and aggregation strip the toe holds used for re-identification. Naturally, legal protections of data are still useful as well.

In addition to the old tools, there are new statistical techniques, like differential privacy, that can be used in some cases to destroy identifiable information in data while preserving preselected characteristics within specified limits. However, differential privacy is only feasible when preparing data for a narrowly specified analysis. It cannot support exploratory analyses. It is also an open question how well differential privacy can protect time-series data, and time-correlated patterns in electricity demand data are a unique feature of smart meter data and of particular interest to researchers.

Most examples of re-identification of data revolve around personal data. For example, public voting records, gender, birth date, and zip code are sufficient to uniquely identify the majority of Americans. If we postulate that a data set consists only of time-series meter readings and zip code level location information, what types of external data might be useful in re-identifying the data? How realistic (or unrealistic) is it to imagine public samples of time series data that would achieve a confident match with de-identified meter data? As a corollary, if we are given only meter data, how much can we learn about the residents that produced the patterns in the data? It is easy to think of examples where simple features will reveal important details — imagine a criminal's alibi ruined by meter data that shows that he was not actually home watching TV at the time of a crime. However, it is hard to generalize the heroic assumptions of hypothetical examples into a coherent model of privacy risks.

We postulate that truly malicious uses of meter data will be rare and likely to require other forms of personal information or physical access additional to the meter data. For example a would-be robber using meter data to guess when someone will be home will

also check for lights on, a car in the driveway, and other signs of life. The most likely violations of privacy using meter data fall into the category of commercial data mining. Data mining techniques are being applied to virtually all forms of online activity and are not necessarily unacceptable. Even if they are uncomfortable with them, many people passively accept such practices. However, given the public-interest nature of smart meters, it would seem that special protections should be extended to customers. Some observers suggest that people will grow more accustomed to their online lives being public over time. If this happens, will it be acceptable to lower privacy standards on meter data?

One area where protecting the identity of customers is potentially counterproductive is in the area of program targeting. If it is possible to determine with precision who will benefit the most from public-interest programs and thereby contribute the most to program goals, why should we dull such insights with data aggregation? Should public-interest programs be allowed to perform precision targeting while service providers and other commercial businesses are not allowed to do the same? A potential solution to this question would be to allow customers to opt into receiving targeted messages or requesting a targeted analysis of their data, but it is unlikely that very many would proactively decide to do so without significant incentives.

Finally, we turn to the question of how the potential benefits of meter data can be achieved. We have suggested that differentiated access to data and the ability for stakeholders to delegate analyses to more trusted entities can jointly protect privacy while preserving flexibility in the types of information available to stakeholders. This flexibility is important because the planned long-term trajectory of the grid (low carbon, renewable, efficient, and with more flexible demand) will require creativity and innovation. It is hard to know in advance who will innovate and how, so it would be unwise to artificially constrain the explorations of stakeholders, provided customer privacy can be protected.

There are many details to consider in a system of delegated analysis with many different types of stakeholders. Public samples of data could be made available to facilitate the development of algorithms. For some stakeholders, routine data metrics and aggregation methods could be largely pre-computed and pre-approved for use. More complex algorithms could be developed in a collaborative relationship with the trusted delegate, subject to review by data privacy experts, or through an adversarial proceeding that would allow public comments and objections to be raised.

Another potential benefit of delegated analysis would be the ability for regulators to observe and learn from the details of uses of the data. Similar in spirit to the basic bargain inherent in a patent application, data requests, algorithms, and returned data could be held for several months or a year before public disclosure to allow innovators some time to benefit from their requests before their competitors learn what they have done. The eventual public disclosure of requests would ensure that stakeholders carefully consider their uses of data. Advocates of the public interest would be able to examine their activity and recommend changes, and customers would be able to learn in detail how their data is being used, possibly increasing their trust in the system.

5.6.1 Implementation and operation

Even if one is supportive of the concept of differentiated access with delegated analysis, several practical concerns remain. How much would such a system cost to setup and maintain? How would it be funded? Who would operate it? This is not the appropriate forum to delve into these issues in detail, but a few clarifying thoughts should help readers judge the feasibility of such a system.

By the standards of modern data centers, which are designed to stream high definition videos, serve millions of web pages, and record and analyze the shopping and web browsing habits of millions of people, the data transfer and storage requirements of smart meter data are modest. A year's worth of 15-minute interval data for 1 million customers is approximately 35 billion observations. With each observation stored comfortably in 2 bytes, a full year's worth of data from 1 million smart meters would occupy 70GB of hard drive space, uncompressed. At the time of this writing, a 4000GB network storage class hard drive, capable of storing a year's worth of data for 56 million people, sells for under $200. Similar arguments apply to the affordability of the network bandwidth and computational power required to transfer and analyze meter data. Hardware and data center bandwidth and hosting costs for aggregated should be affordable.

The most significant project costs would likely be for the development and maintenance of the software systems and the associated professional services. These would, in turn, strongly depend on the features desired and how actively the system is used and updated. Several mature open source software projects, like Big Table [19], Hadoop [116], Hive [112], and Dremel [74] already support large scale data storage and processing. The free and open statistical and scientific computing libraries in Python [49] and R [111] are among the most sophisticated available at any price. If open source software is adopted, the bulk of the development costs would go to system design and integration.

Assuming that the system ultimately costs a one to two million dollars to build and several hundred thousand dollars a year to maintain[15], who should bear the costs? One option would be to pay for the system using ratepayer funds, possibly by drawing on the program planning, execution, and evaluation budgets that such a system would help to reduce. In California, program budgets run in the hundreds of millions of dollars a year, so even minor improvements in efficacy would pay for a delegated analysis system while money to spare. Another option would be to charge user fees to the stakeholders using the system. Because private industry stands to benefit from access to the data, they could likely be persuaded to contribute user fees. The main concerns raised by a dependence on user fees would be whether they will be consistent over time and whether the system would become dependent on larger users, who would then have significant leverage over the future development of the system.

A final issue to discuss is who or what would serve as the trusted delegate for delegated

[15]The core functionality could certainly be implemented for cheaper, and there are many ways the budgets could grow beyond these numbers, but they should be the right order of magnitude for a well-scoped, -built and -maintained system.

analysis. One option would be to employ one or more trusted data savvy professionals to help stakeholders develop and run analyses. Depending on the originality of each request and the level of interest in the service, these people could become overwhelmed with requests. It would be technically feasible to make most or all delegated data analysis self-service. Software that provides an execution environment with restricted privileges is called a *sandbox*. Examples of sandbox environments include smart phones, where applications must request and be granted access to phone features before using them, and web browsers, which run JavaScript and Flash code downloaded from the internet while preventing that code from accessing system resources like the local hard drive. Similarly, a data sandbox could provide pre-approved analytical capabilities that untrusted stakeholders would be allowed to use without gaining access to any sensitive data.

5.7 Conclusions

A wide variety of potential benefits beyond avoiding manual meter reading and enabling new rate designs come with the installation of smart meter infrastructure. In particular, spatial and temporal patterns in demand data can be used to improve grid planning and operations; inform energy efficiency and demand response program planning, implementation, and evaluation; and help efficiency and solar service providers shape their products and find customers who will benefit from their services. However, the timing and magnitude of customers' energy consumption can also reveal private information about them. This poses a unique challenge to regulators. They must figure out how to encourage innovative uses of meter data in support of public-interest goals while minimizing the potential for misuse of disclosed data.

In California, the use of identifiable customer data has been restricted to utilities and their selected third parties in the support of system, grid, or operational needs, and the implementation of demand response, energy management, or energy efficiency programs. Utilities and selected third parties are prohibited from selling customer data, which reflects an awareness that there is money to be made using customer data for commercial purposes that are not in the public interest. Yet, commercial third parties with data expertise, who could not legally buy access to meter data, are paid to work with it while researchers, municipalities, service providers, and other stakeholders who play important roles in delivering the benefits of public-interest programs are not typically granted access to the data.

We have argued that utilities, contractors, regulators, researchers, commercial businesses, and customers should be granted *differentiated access* to smart meter data, with different levels of detail available according to the needs, motivation, capabilities, and trustworthiness of each stakeholder. However, best-practice data privacy protections, including legal protections, respect for customer preferences, data minimization, de-identification, and aggregation should be applied consistently across all stakeholders.

Achieving a low-carbon, efficient, flexible, reliable, and affordable grid will require

learning by doing and many innovative changes to business as usual. Demand data should not be so restricted that it prevents stakeholders from playing constructive roles in meeting these challenges. We observe that the outputs of energy data analyses can be designed to be less revealing than their inputs. It follows that a stakeholder lacking sufficient privileges under differentiated access to perform an analysis itself could engage a trusted delegate to access the sensitive data and perform the analysis, returning only the anonymized results. This approach, which we call *delegated analysis*, holds a great deal of promise for offering ongoing flexibility of analysis while preserving privacy protections. Other approaches, like pre-computed data aggregations or differential privacy, are tailored to the data requirements of a particular analysis and therefore cannot support innovative analyses of meter data in the same way. However, if desired, the trusted delegate could enforce aggregation rules or add noise to data to achieve compatibility with these other techniques.

Finally, we have provided many examples of the types of beneficial analysis that could be performed by utilizing delegated analysis. The examples provided are not meant to be an exhaustive list, but they cover applications relevant to a range of stakeholder types and should provide ample evidence of the flexibility of the delegated analysis approach.

Smart meter data provides a view of the spatial and temporal patterns of energy consumption in unprecedented detail at a time when demand-side program are growing in importance. Information about individual customer contributions to demand can be used to improve grid operations and the outcomes of energy efficiency, demand response, and grid planning. Regulators have the challenge of ensuring that meter data is used to its full potential while maintaining the privacy of utility customers. This will require consistently applied restrictions on access to meter data that preserve flexibility in its use over time. This paper provides a practical approach to data privacy and analysis that strikes this balance.

Chapter 6

Concluding remarks

In most service territories, the primary argument for installing smart meter infrastructure is the support of utility operations. Smart meters reduce meter reading costs, provide operational diagnostics, inform long-term planning, and enable the deployment of time-of-use and real-time rate plans. These operational benefits alone justify the expense of smart meter infrastructure installation and maintenance. However, the benefits achievable through creative applications of the empirical energy consumption data smart meters provide could easily surpass the operational benefits.

It has long been understood that energy consumption tends to follow basic patterns. For example, the hotter the climate, the more air conditioning usage; the wealthier the consumer, the more they consume. Traditional methods of characterizing and predicting consumption in support of energy efficiency and demand response programs use these insights, coupled with consumer marketing and neighborhood aggregate demographic data, to make predictions about appliance ownership and levels of consumption in different households.

Those predictions are used to allocate finite energy efficiency and demand response program resources. However, results from Chapters 2 and 3 clearly demonstrate that the variability within well defined categories, like climate zones or income level, tends to exceed the variability across categories. This dissertation presents a variety of methods, ranging from simple to complex, for extracting consumption features from the meter data of individual households. These techniques capture the variability within the traditional categories of consumption and can support the far more accurate targeting of energy efficiency and demand response programs than current best practices allow.

This dissertation has also demonstrated that distributions drawn from individual household features can help to improve characterizations of the nature of the building stock. Chapters 3 and 4 show that features derived from meter data can be used to estimate physical characteristics of homes and to infer information about lifestyle choices, appliance ownership, and patterns of occupancy. These estimates collectively comprise a data resource that can be used to train and test predictive algorithms, guide strategic utility planning, and inform academic research into the nature of energy consumption.

Potential applications of meter data are not limited to beneficial uses. Chapter 5 develops a framework for understanding potential benefits and harms that are inherent in the availability of meter data, with a focus on the benefits that can be realized by public interest energy efficiency and demand response programs and harms that may include the loss of privacy for utility customers. Through a variety of examples, this chapter illustrates the feasibility of beneficial uses and the nature of realistic privacy risks. It concludes with a description of a system that would grant differentiated access to data according to the risks posed by each stakeholder, with some forms of analysis remaining accessible to less trusted stakeholders through data access and analysis delegated to a trusted party. This system demonstrates that it is possible to support innovative uses of meter data while protecting customer data from unwelcome uses.

The burden of striking the balance between the beneficial uses of meter data (known and unknown) and protecting the rights of customers ultimately falls to utility regulators.

The decisions they make will determine the ultimate fate of the techniques developed in this dissertation and related fields of research. With their support of ongoing innovation and protection of customer rights, smart meter data could fuel a set of powerful tools that facilitate the transition to clean, flexible, reliable and affordable grid.

6.1 Future and related work

After more than 30 years of innovation in energy efficiency programs, smart meter data has recently emerged as a powerful new tool whose beneficial impact should be felt by virtually every aspect of program design, implementation and evaluation. These benefits come at a time of significant pressures for innovation and change within the utility industry. Because it characterizes the consequences of the choices made by their customers, meter data is fertile ground for innovation in most domains of utility activity. This section outlines several potential areas of future work involving meter data. They are either natural extensions of the work presented in this dissertation or similar enough in spirit to be relevant to the discussion of related work.

6.1.1 Testing targeting outcomes of real programs

This dissertation includes many examples of *empirical targeting* criteria expected to improve the outcomes of efficiency and demand response programs by locating the specific households most likely to benefit from program offers. However, these expectations have not yet been tested under real-world conditions. A proper verification of empirical targeting's potential is a logical next step.

The gold standard for this work is a randomized controlled trial, which would require cooperation and support from a program administrator. The basic structure of the study would be to divide all households into four groups with similar attributes. For a large sample, this can be accomplished by randomly assigning each household to one of four groups. One group should be designated the control group and not contacted or encouraged to participate in the program in any way. Within each of the remaining groups, the same number of households should be selected to receive information about the program being studied. Households from one group should be selected at random; households from the another group should be targeted using current best practices; and household from the last group should be selected using empirical targeting. To the extent possible, all other details — the timing of messaging, measure implementation timing (if applicable), and the characteristics of any professionals involved — should be identical across groups, unless the differences are somehow part of the targeting method being studied. As a practical matter, the number of households contacted could be determined by the program's implementation budget.

Because of the design of the trial, differences in outcomes between groups will be attributable to the differences in targeting methods. The participation rates and program

relevant outcomes from each group, i.e., kWh saved, appliances installed, demand response curtailment observed, etc., could be compared to the others, including the control group, and used to quantify the efficacy of the different targeting approaches. To isolate the effect of targeting alone, the same message could be provided to all selected households. To capture the full potential of household-specific information, a more tailored message (which could influence participation rates) could be presented to the empirically targeted households. If desirable, a larger number of groups could be used to test additional program targeting permutation.

The biggest drawbacks of randomized controlled trials are that they require one group to be left out of participation in a beneficial program and they require discipline, time, and effort to conduct. Because program budgets cannot support serving all households, the first concern is relatively minor — no matter what, some households will not be able to participate. The second concern is a significant obstacle. From planning through implementation and evaluation, energy efficiency program cycles typically span several years. The resulting delay in obtaining final results is therefore inevitable for a complete study. This could have a chilling effect on the rate of innovation in program targeting.

However, there are a few alternatives that should provide insights into the efficacy of program targeting in a more timely manner. First, trials can be designed to support ongoing monitoring and the determination of tentative results. Second, when the details of a program's implementation allow, the time required to get results can be reduced through staged trials — with a new set of participants for each of several phases of program implementation. Third and most immediately, past program results can be used as natural experiments in program outcomes. Ignoring known outcomes, control and participation groups could be assigned, with different targeting criteria used to select households within each. The known program outcomes for the selected households could then be compared to determine how reliably the targeting criteria picked participating households and quantify differences in program outcomes.

6.1.2 Improving program evaluation

The determination of outcomes from energy efficiency and demand response programs is a costly and resource intensive process. The methods used are heterogeneous and subject to reasonable disagreements over their findings. There is currently significant interest in new evaluation methods that make use of smart meter data, particularly methods that take advantage of randomized controlled trials. A full treatment of this topic is outside the scope of this document. However, there is one approach to evaluation that deserves more attention: the use of meter data to attribute changes in consumption to specific interventions through the temporal patterns of the observed changes.

One of the most difficult aspects of program evaluations is that the conditions that determine energy consumption in households are always changing. Evaluations of interventions necessarily involve the comparison of consumption from a time before and a time after the intervention. In the intervening span of time, other aspects of household energy

use can change and confound estimates of program outcomes. Some programs resolve this difficulty by relying on engineering estimates of intervention impacts, but this approach relies on assumptions that may not hold and eliminates the opportunity to observe and learn from unexpected outcomes. Some programs rely on statistical tools to try to isolate program effects, but these tools come with their own assumptions and reasonable people can disagree on both methods and assumptions. Some programs do not control for outside changes at all. Instead they rely on the assumption that outside changes will cancel out in a large sample and report aggregate, rather than individual outcomes. These assumptions may not hold in the case of universal trends and some evaluations require estimates of individual outcomes.

Because smart meter data resolves patterns of consumption in time, a new option for isolating program impacts is available. Every measure category can be expected to impact patterns of consumption differently. For example, a more efficient refrigerator will lower consumption by a (nearly) fixed amount around the clock; an air conditioning tune-up will only reduce consumption while cooling is in use during hours of occupancy on hot days; and a lighting retrofit will tend to reduce consumption more during the night than during the day.

Both empirical and engineering models of the patterns of expected measure impacts can be constructed and used to identify patterns in the changes in consumption seen in data from before and after an intervention. Research into such methods pose exciting technical challenges and could prove to be of significant social value.

6.1.3 Holistic grid planning

Because they involve different domains of knowledge and superficially different goals, planning processes for grid capacity, renewable energy deployment, energy efficiency programs, and demand response programs tend to be conducted independently. However these programs have the potential to be either mutually beneficial or mutually antagonistic. The strategic integration of planning and objectives across these interests would produce benefits greater than the sum of benefits achieved individually. For example, infrastructure constraints like limits on transmission capacity are more binding in some locations and times than others. Finite efficiency and demand response program resources should be focused on geographic regions and times of use where they can best help to alleviate these constraints. For example air conditioning efficiency and control measures should be considered more valuable and implemented with greater urgency in regions that contribute most to peak cooling demands along congested transmission corridors.

An extension of the above example involves the integration of renewable resources, which are not direct replacements for conventional power plants. They are inherently spatially distributed and their production is naturally intermittent. The wind doesn't always blow and the sun doesn't always shine. This means that binding constraints on a highly renewable grid will be more widely distributed in time and space. However, wind and solar resources do have prevailing characteristics in different locations and are subject

to forecasts that grow more and more accurate as their prediction window approaches. Infrastructure planning for a renewable grid must be responsive to these new challenges, but energy efficiency and demand response programs can also help to manage the timing and reduce the severity of binding constraints. This is particularly true when the efficiency and demand response resources have characteristics that are well matched to the prevailing patterns of renewable energy intermittency. For example, solar production reliably ramps down in the evening, potentially creating a shortfall in available energy. Improvements to the efficiency and control of end uses that ramp up in the evenings — like lighting and electric heating — can help offset these effects.

Finally, it is worth considering the emissions impacts of avoided consumption. A change in one household's consumption will reduce demand for power in a manner that most directly impacts the marginal power producer. The marginal producer will thus determine the emissions impact of that change. The marginal producer in a service territory can change from hour to hour, with some consistency in general patterns over time. The timing of savings achieved through energy efficiency and demand response interventions will thus determine their emissions impacts. Assuming a baseload coal plant is on the margin overnight, measures that reduce overnight emissions, i.e., more efficient lighting, refrigeration, and electric heating and reduced standby power consumption of electronics, will have a greater mitigation impact than those that reduce mid-day consumption.

These examples point toward grid management benefits achievable through the strategic use of energy efficiency and demand response measures once the temporal patterns and reliability of their impacts are more precisely known. The techniques for improving attribution of measure impact discussed in Section 6.1.2 are relevant to this application as well.

6.1.4 Ongoing research opportunities

As early work using a new source of data, this dissertation raises more questions than it answers. It focuses substantial attention on techniques anticipated to be of immediate practical value, however much of the work originated with questions that were more theoretical or open ended in nature. In that spirit, this final section offers a selection of unresolved research questions about applications of smart meter data raised by this work.

- Efficiency program planners rely heavily on estimates of appliance ownership and end use breakdowns within the service territories they work in. Traditionally, this information is obtained using labor intensive and expensive survey methods, which must be repeated on a regular basis. To what extent can smart meter data be used to the reproduce appliance saturation and end use breakdowns currently captured through survey instruments? What derived metrics and related analytical techniques are the most appropriate to this type of stock characterization?

- How different are the details of modeling outcomes and estimated stock characteristics when using meter data from different service territories? What differences

are driven by climate? Is there evidence of behavioral and cultural similarities or differences across service territories? Is there evidence of regional differences in construction details, like insulation quality, air tightness, and home size? To what extent are lessons learned or distributions generated in one territory applicable to others?

- Targeting grows more fruitful as the differences between households in a sample become more extreme. Borrowing an example from the study of income inequality, Gini coefficients could potentially be used to quantify the disparity between high and low values of useful metrics within populations. To what extent would these coefficients capture real world targeting potential? Could they be used to identify 'good enough' targeting methods that are either more computationally efficient or more protective of customer privacy than the absolute best methods?

- How might change detection algorithms be applied to meter data model outcomes? Could they detect program interventions? Could they identify lifestyle changes, new appliances, or the arrival of new residents? The ability to reliably identify any of these changes could be used to expand our collective understanding of the drivers of energy consumption and the behavioral factors that differentiate the consumption of different households.

- It is anticipated that the disaggregation of end uses from whole home meter data will help to locate waste in homes and communicate with residents about their potential savings. There is already significant attention focused on applicable algorithms and suitable metrics of their performance. However, the research questions in the field of disaggregation could be tweaked to better support the needs of program administrators: How well can disaggregation approaches estimate the presence or absence (as opposed to the energy consumption) of particular end uses or equipment? Are disaggregation algorithms able to detect and attribute the changes in consumption that occur as the result of energy efficiency or demand response interventions? Assuming smart meter infrastructure will remain bandwidth constrained, how much data sampled with what frequency is needed to address the preceding questions?

- Meter data includes consumption from both deterministic systems and the behaviors of occupants. How can these two categories of consumption best be disaggregated? What insights about human behavior can be developed using meter data? How can these insights be applied to improve the outcomes of energy efficiency and demand response programs?

- An important tool of energy efficiency and demand response marketing is market segmentation, where customers are divided into groups with similar traits. Different marketing strategies inspired by relevant traits can then be developed for each group. What smart-meter-data-derived features can be used to segment customers? How

well do those segments perform compared to conventional psychographic segments? How can statistical classification and clustering methods be adapted to produce segments useful to program planners?

- One potential application of meter data is to educate and inform customers. What energy concepts resonate with customers? What methods of presentation and data interaction do customers respond to? What messages based on meter data contents inspire action on the part of customers? How well can customers recognize and accurately assess their own energy waste when they are supported by smart meter data visualization and analysis tools?

- How well do the methods and lessons learned using residential data translate to commercial customers? What new tools and performance metrics are required to address the commercial market? How important are the differences between categories of commercial activity, i.e., retail, office, health care, etc.? How should lessons drawn from commercial meter data analysis be informed by the high levels of automation in commercial buildings?

Despite years of intensive study, it is clear that the methods and insights developed in this dissertation have only scratched the surface of potential applications of meter data. The untapped potential of this new resource is encouraging given the magnitude of the challenges inherent in adapting the grid to the needs of the 21st century. The timely transition to a clean, flexible, affordable, and reliable grid will require creative and ground-breaking work that draws upon domain knowledge from utilities, regulators, and academic disciplines — including building science, product design, electric power systems, public policy, data mining, energy economics, psychology, and behavioral sciences. The stakes are high and time is limited. It is time to put smart meter data to work.

Bibliography

[1] Joana M. Abreu, Francisco Camara Pereira, and Paulo Ferrao. Using pattern recognition to identify habitual behavior in residential electricity consumption. *Energy and buildings*, 49:479–487, 2012.

[2] Adrian Albert and Ram Rajagopal. Smart Meter Driven Segmentation: What Your Consumption Says About You. *IEEE Transactions on Power Systems*, 28(4):4019–4030, November 2013.

[3] Hunt Allcott. Social norms and energy conservation. *Journal of Public Economics*, 95(9):1082–1095, 2011.

[4] Hunt Allcott and Sendhil Mullainathan. Behavior and Energy Policy. *Science*, 327(5970):1204–1205, March 2010. PMID: 20203035.

[5] K. Carrie Armel, Abhay Gupta, Gireesh Shrimali, and Adrian Albert. Is disaggregation the holy grail of energy efficiency? The case of electricity. *Energy policy*, 2012.

[6] Maximilian Auffhammer, Carl Blumstein, and Meredith Fowlie. Demand-side management and energy efficiency revisited. 2007.

[7] Peder Bacher and Henrik Madsen. Identifying suitable models for the heat dynamics of buildings. *Energy and buildings*, 43:1511–1522, 2011. 7.

[8] Galen L. Barbose, Charles A. Goldman, Ian M. Hoffman, and Megan Billingsley. The Future of Utility Customer-Funded Energy Efficiency Programs in the United States: Projected Spending and Savings to 2025. Technical Report 5803E, LBNL, 2013.

[9] Mangesh Basarkar, Xiufeng Pang, Liping Wang, Philip Haves, and Tianzhen Hong. Modeling and Simulation Of HVAC Faults in EnergyPlus. 2011.

[10] Christian Beckel, Leyna Sadamori, and Silvia Santini. Automatic socio-economic classification of households using electricity consumption data. pages 75–86. ACM, 2013.

[11] Benjamin J. Birt, Guy R. Newsham, Ian Beausoleil-Morrison, et al. Disaggregating categories of electrical energy end-use from whole-house hourly data. *Energy and buildings*, 50:93–102, 2012.

[12] Blumstein. Overcoming Social and Institutional Barriers to Energy Conservation. *Energy*, 5:355, 1980. 4.

[13] Carl Blumstein, C. Bart McGuire, and Susan Buller. Distribution of the intensity of electricity use in commercial premises appears to be highly skewed. *Energy and buildings*, 10:151–153, 1987. 2.

[14] Hai Brenner and Kobbi Nissim. Impossibility of differentially private universally optimal mechanisms. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, page 71–80, 2010.

[15] Marilyn A. Brown, Mark Levine, Jonathan Koomey, Lynn Price, and Nathan Martin. Scenarios of U.S. Carbon Reductions: Potential Impacts of Energy-Efficient and Low-Carbon Technologies by 2010 and Beyond. Technical report, 1997.

[16] Rebecca Brown. Modeled vs. Actual Energy Savings for Energy Upgrade California Home Retrofits. Technical report, BKi, PIER Project PIR-08-018, 2012.

[17] Kenneth P. Burnham. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer, New York, 2nd ed edition, 2002.

[18] Ann Cavoukian and Jules Polonetsky. Privacy by Design and Third Party Access to Customer Energy Usage Data. Technical report, January 2013.

[19] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, et al. Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems (TOCS)*, 26(2):4, 2008.

[20] Jeremy Cherfas. Skeptics and visionaries examine energy saving. *Science (Washington, DC);(United States)*, 251(4990), 1991.

[21] Consortium for Energy Efficiency. 2013 Behavior Program Summary - Public Version. March 2013.

[22] D. L. Costa and M. E. Kahn. Energy conservation" nudges" and environmentalist ideology: Evidence from a randomized residential electricity field experiment. Technical report, National Bureau of Economic Research, 2010.

[23] S. Darby. The Effectiveness of Feedback on Energy Consumption. *A Review for DEFRA of the Literature on Metering, Billing and Direct Displays*, April 2006.

[24] RJ de Dear and GS Brager. Developing an adaptive model of thermal comfort and preference. *ASHRAE Transactions*, 104:145–167, 1998. 1.

[25] Thomas Dietz, Gerald T. Gardner, Jonathan Gilligan, Paul C. Stern, and Michael P. Vandenbergh. Household actions can provide a behavioral wedge to rapidly reduce U.S. carbon emissions. *PNAS*, 106:18452–18456, 2009. 44.

[26] Cynthia Dwork. A firm foundation for private data analysis. *Communications of the ACM*, 54(1):86–95, 2011.

[27] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, page 265–284. Springer, 2006.

[28] Richard E. Edwards, Joshua New, and Lynne E. Parker. Predicting future hourly residential electrical consumption: A machine learning case study. *Energy and Buildings*, 49:591–603, June 2012.

[29] Costas Efthymiou and Georgios Kalogridis. Smart grid privacy via anonymization of smart metering data. In *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on*, page 238–243, 2010.

[30] EIA. 2003 Commercial Buildings Energy Consumption Survey. Technical report, EIA, Washington, DC, 2006.

[31] EIA. 2009 Residential Energy Consumption Survey. Technical report, 2011.

[32] EPA. Portfolio Manager Technical Reference: Energy STAR Score, July 2013.

[33] Margaret F. Fels. PRISM: an introduction. *Energy and Buildings*, 9(1):5–18, 1986.

[34] Hilary Forster, Patrick Wallace, and Nick Dahlberg. 2012 State of the Efficiency Program Industry Report. Technical report, March 2013.

[35] Jon Froehlich, Eric Larson, Sidhant Gupta, et al. Disaggregated End-Use Energy Sensing for the Smart Grid. *IEEE Pervasive Computing*, 10:28–39, 2010. 1.

[36] Zachary M. Gill, Michael J. Tierney, Ian M. Pegg, and Neil Allan. Low-energy dwellings: the contribution of behaviours to actual performance. *Building Research & Information*, 38(5):491–508, 2010.

[37] K. Gillingham, R. G. Newell, and K. Palmer. Energy efficiency economics and policy. *Annual Review of Resource Economics*, page 597–619, 2009. 1.

[38] David Goldstein. Extreme Efficiency: How Far Can We Go If We Really Need To? ACEEE, 2008.

[39] Philippe Golle. Revisiting the uniqueness of simple demographics in the US population. In *Proceedings of the 5th ACM workshop on Privacy in electronic society*, page 77–80, 2006.

[40] J. S. Haberl and S. Thamilseran. The great energy predictor shootout II: Measuring retrofit savings. *ASHRAE journal*, 40:49–56, 1998. 1.

[41] G.W. Hart. Residential energy monitoring and computerized surveillance via utility power flows. *IEEE Technology and Society Magazine*, 8(2):12–16, 1989.

[42] David J. Hess and Jonathan S. Coley. Wireless smart meters and public acceptance: The environment, limited choices, and precautionary politics. *Public Understanding of Science*, 2012.

[43] Karin Hieta, Valerie Kao, and Thomas Roberts. Case Study of Smart Meter System Deployment: Recommendations for Ensuring Ratepayer Benefits. Technical report, Division of Ratepayer Advocates, San Francisco, CA, March 2012.

[44] Tyler Hoyt, Kwang Ho Lee, Hui Zhang, Edward Arens, and Tom Webster. Energy savings from extended air temperature setpoints and reductions in room air mixing. 2009.

[45] David Hsu. Characterizing Energy Use in New York City Commercial and Multifamily Buildings. *ACEEE Summer Study on Energy Efficiency in Buildings*, 2012.

[46] Institute for Electric Efficiency. Utility-Scale Smart Meter Deployments, Plans & Proposals. Technical report, Edison Foundation, May 2012.

[47] Itron. California Commercial End-Use Survey. Technical report, CEC, 2006.

[48] K. B. Janda. Buildings don't use energy: people do. *Architectural Science Review*, 54:15–22, 2010. 1.

[49] Eric Jones, Travis Oliphant, and Pearu Peterson. SciPy: Open source scientific tools for Python. *http://www. scipy. org/*, 2001.

[50] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, page 263–291, 1979.

[51] KEMA. 2009 California Residential Appliance Saturation Study. Technical report, California Energy Commission, October 2010.

[52] John Kissock, FJeff S. Haberl, and David E. Claridge. Inverse Modeling toolkit: Numerical Algorithms. *Ashrae Transactions: Symposia*, 2003.

[53] Robert Kohlenberg, Thomas Phillips, and William Proctor. A Behavioral Analysis Of Peaking In Residential Electrical-energy Consumers. *Journal of Applied Behavior Analysis*, 9(1):13–18, 1976.

[54] J. Z. Kolter, S. Batra, and A. Y. Ng. Energy disaggregation via discriminative sparse coding. *Neural Information Processing Systems*, 2010.

[55] J. Z. Kolter and J. Ferreira Jr. A Large-Scale Study on Predicting and Contextualizing Building Energy Usage. AAAI, 2011.

[56] J. Zico Kolter and Matthew J. Johnson. REDD: A Public Data Set for Energy Disaggregation Research. SustKDD, 2011.

[57] J. F. Kreider and J. S. Haberl. Predicting hourly building energy use: the great energy predictor shootout- overview and discussion of results. *ASHRAE Transactions*, 100:1104–1118, 1994. 2.

[58] Klaus Kursawe, George Danezis, and Markulf Kohlweiss. Privacy-friendly aggregation for the smart-grid. In *Privacy Enhancing Technologies*, page 175–191, 2011.

[59] J Kwac, J Flora, and Ram Rajagopal. Household Energy Consumption Segmentation using Hourly Data. *Smart Grid, IEEE Transactions*, 2013.

[60] Jungsuk Kwac. *akmeans: Adaptive Kmeans algorithm based on threshold*. 2013. R package version 1.0.

[61] Audrey Lee. Energy data center briefing paper. Technical report, California Public Utilities Commission, San Francisco, CA, September 2012.

[62] Steven B. Leeb, Steven R. Shaw, and Jr. James L. Kirtley. Transient Event Detection in Spectral Envelope Estimates for Nonintrusive Load Monitoring. *IEEE Transactions on Power Delivery*, 10:1200–1210, 1995. 3.

[63] M. Levine, D. Urge-Vorsatz, K. Blok, et al. Residential and Commercial Buildings. Climate Change 2007; Mitigation. Contribution of Working Group III to the Fourth Assessment Report of the IPCC. *Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA*, 2007.

[64] Arik Levinson. California Energy Efficiency: Lessons for the Rest of the World, or Not? Working Paper 19123, National Bureau of Economic Research, June 2013.

[65] Eckhard Limpert, Werner A. Stahel, and Markus Abbt. Log-normal distributions across the sciences: keys and clues. *BioScience*, 51(5):341–352, 2001.

[66] Lovins. The Super-efficient Passive Building Frontier. *ASHRAE journal*, 37:79, 1995. 6.

[67] L. Lutzenhiser, L. Cesafsky, H. Chappells, et al. Behavioral assumptions underlying California residential sector energy efficiency programs. *Prepared for the California Institute for Energy and Environment Behavior and Energy Program*, 2009.

[68] Loren Lutzenhiser. Social And Behavioral Aspects of Energy Use. *Annu. Rev. Energy. Environ.*, 1993:247–89, 1993. 18.

[69] M. Dworkin, K. Johnson, D. Kreis, et al. A Regulator's Privacy Guide to Third-Party Data Access for Energy Efficiency. Technical report, State and Local Energy Efficiency Action Network, Lawrence Berkeley National Lab, 2012.

[70] Henrik Madsen and Jan Holst. Estimation of continuous-time models for the heat dynamics of a building. *Energy and Buildings*, 22(1):67–79, 1995.

[71] Patrick McDaniel and Stephen McLaughlin. Security and privacy challenges in the smart grid. *Security & Privacy, IEEE*, 7(3):75–77, 2009.

[72] Eoghan McKenna, Ian Richardson, and Murray Thomson. Smart meter data: Balancing consumer privacy concerns with legitimate applications. *Energy Policy*, 41:807–814, 2012.

[73] Fintan McLoughlin, Aidan Duffy, and Michael Conlon. Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: An Irish case study. *Energy and Buildings*, 48:240–248, May 2012.

[74] Sergey Melnik, Andrey Gubarev, Jing Jing Long, et al. Dremel: interactive analysis of web-scale datasets. *Proceedings of the VLDB Endowment*, 3(1-2):330–339, 2010.

[75] E. Mills, P. Mathew, and M. A. Piette. Action-Oriented Benchmarking: Concepts and Tools. *Energy Engineering*, 105:21–40, 2008. 4.

[76] Evan Mills. Building Commissioning: A Golden Opportunity for Reducing Energy Costs and Greenhouse-gas Emissions. Technical report, LBNL, 2009.

[77] Cynthia Mitchell. Stabilizing California's Demand. *Public Utilities Fortnightly*, page 50–58, 2009.

[78] M. Moezzi, L. Lutzenhiser, and J. Woods. Behavioral assumptions in energy efficiency potential studies. *May. Prepared for the California Institute for Energy and Environment (CIEE). Oakland, Calif*, 2009.

[79] R. E. Mortensen and K. P. Haggerty. A stochastic computer model for heating and cooling loads. *Power Systems, IEEE Transactions on*, 3(3):1213–1219, 1988.

[80] Kazunori Nagasawa, Charles R. Upshaw, Joshua D. Rhodes, et al. Data Management For A Large-scale Smart Grid Demonstration Project In Austin, Texas. In *Proceedings of the ASME 6th International Conference on Energy Sustainability*, San Diego, CA, 2012.

[81] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, page 111–125, 2008.

[82] Arvind Narayanan and Vitaly Shmatikov. De-anonymizing social networks. In *Security and Privacy, 2009 30th IEEE Symposium on*, page 173–187, 2009.

[83] Arvind Narayanan and Vitaly Shmatikov. Myths and fallacies of personally identifiable information. *Communications of the ACM*, 53(6):24–26, 2010.

[84] LK Norford, R. H. Socolow, E.S. Hsieh, and G.V. Spadaro. Two-to-One Discrepancy Between Measured and Predicted Performance of a Low-energy Office Building - Insights From a Reconciliation Based on the DOE-2 Model. *Energy and buildings*, 21:121, 1994. 2.

[85] NRC. *Energy Research at DOE: Was It Worth It? Energy Efficiency and Fossil Energy Research 1978 to 2000*. NRC, 2001.

[86] Paul Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, 57:1701, 2010.

[87] Danny S. Parker. Research highlights from a large scale residential monitoring study in a hot climate. *Energy and buildings*, 35:863–876, 2003. 9.

[88] F. J. Peterson, J. E. Patton, M. E. Miller, et al. End-Use Load and Consumer Assessment Program: motivation and overview. *Energy and buildings*, 19:159–166, 1993. 3.

[89] PG&E. Pacific Gas And Electric Company Smart Grid Annual Privacy Report 2012. Technical report, April 2013.

[90] PG&E Corp. SEC 10-K filing, 2009.

[91] Phil Price. Methods for Analyzing Electric Load Shape and its Variability. Technical report, LBNL, Berkeley, 2010. LBNL-3713E.

[92] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2012. ISBN 3-900051-07-0.

[93] A. Rabl. Parameter estimation in buildings: methods for dynamic analysis of measured energy use. *Journal of Solar Energy Engineering*, 110(1):52–66, 1988.

[94] A. Rabl and A. Rialhe. Energy signature models for commercial buildings: test with measured data and interpretation. *Energy and buildings*, 19:143–154, 1992. 2.

[95] Vibhor Rastogi and Suman Nath. Differentially private aggregation of distributed time-series with transformation and encryption. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, page 735–746, 2010.

[96] Howard Reichmuth and Cathy Turner. A Tool for First Views of Building Energy Performance. Technical report, New Buildings Institute, 2011. Grant 83378201.

[97] Richard Rhodes and Beller. The Need for Nuclear Power. *Foreign Affairs*, 79, 2000. 1.

[98] Alfredo Rial and George Danezis. Privacy-preserving smart metering. In *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society*, page 49–60, 2011.

[99] A. H. Rosenfeld and D. Poskanzer. A Graph Is Worth a Thousand Gigawatt-Hours: How California Came to Lead the United States in Energy Efficiency (Innovations Case Narrative: The California Effect). *Innovations: Technology, Governance, Globalization*, 4:57–79, 2009. 4.

[100] Mike Rufo and Fred Coito. California's Secret Energy Surplus: The Potential For Energy Efficiency. Technical report, prepared by Xenergy Inc. for the Energy Foundation and Hewlett Foundations, October 2002.

[101] SCE. Southern California Edison Company's Annual Privacy Report 2012. Technical report, April 2013.

[102] D. Schwartz, B. Fischhoff, T. Krishnamurti, and F. Sowell. The Hawthorne effect and energy awareness. *Proceedings of the National Academy of Sciences*, 110(38):15242–15246, September 2013.

[103] SDG&E. Annual Privacy Report of SDG&E 2012. Technical report, May 2013.

[104] C Seligman, LJ Becker, and JM Darley. Encouraging residential energy conservation through feedback. *Energy: Psychological Perspectives*, page 93, 1981.

[105] C. Seligman, J. M. Darley, and L. J. Becker. Behavioral approaches to residential energy conservation. *Energy and Buildings*, 1:325–337, 1978. 3.

[106] Elaine Shi, T.-H. Hubert Chan, Eleanor G. Rieffel, Richard Chow, and Dawn Song. Privacy-Preserving Aggregation of Time-Series Data. In *NDSS*, 2011.

[107] Olivier Sidler, B. Lebot, and L. Pagliano. Electricity Demand in European Households: Major Findings from an Extensive End-Use Metering Project in Four Individual Countries. In *American Council for Energy-Efficient Economy*, 2002.

[108] R. H. Socolow. The Twin Rivers program on energy conservation in housing: Highlights and conclusions. *Energy and buildings*, 1:207–242, 1978. 3.

[109] R. C. Sonderegger. Movers and stayers: the resident's contribution to variation across houses in energy consumption for space heating. *Energy and buildings*, 1:313–324, 1978. 3.

[110] Latanya Sweeney. Uniqueness of simple demographics in the US population. *LIDAP-WP4. Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA*, 2000.

[111] R. Developement Core Team. R: A language and environment for statistical computing. Technical report, ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria, 2013. url: http://www. R-project. org, 2005.

[112] Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, et al. Hive-a petabyte scale data warehouse using hadoop. In *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*, page 996–1005, 2010.

[113] Cathy Turner. Energy Performance of LEED NC Buildings. Technical report, New Buildings Institute/US Green Building Council, 2007.

[114] U. S. Census Bureau. *American Community Survey*. U.S. Census Bureau, Washington, DC, 2009.

[115] Christopher Warner. Working Group Report on Rulemaking 08-12-009 Phase III Energy Data Center. Technical report, CPUC, Sam Francisco, July 2013.

[116] Tom White. *Hadoop: the definitive guide*. O'Reilly, 2012.

[117] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009.

[118] Michael Zeifman and Kurt Roth. Nonintrusive appliance load monitoring: Review and outlook. *Consumer Electronics, IEEE Transactions on*, 57(1):76–84, 2011.