Data-Driven Analysis of Time of Day Pricing for Residential Consumers

by

Saba Nejad

Submitted to the Institute for Data, Systems, and Society and the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of
Master of Science in Technology and Policy and Master of Science in Computer Science and Engineering
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
May 2022

Author .....
Institute for Data, Systems, and Society and the Department of Electrical Engineering and Computer Science
May 23, 2022

Certified by .....
Munther A. Dahleh
Director, Institute for Data, Systems, and Society
William A. Coolidge Professor, Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by .....
Noelle Eckley Selin
Professor, Institute for Data, Systems, and Society and Department of Earth, Atmospheric and Planetary Sciences
Director, Technology and Policy Program

Accepted by .....
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Data-Driven Analysis of Time of Day Pricing for Residential Consumers

by

Saba Nejad

Submitted to the Institute for Data, Systems, and Society and the Department of Electrical Engineering and Computer Science
on May 23, 2022, in partial fulfillment of the
requirements for the degree of
Master of Science in Technology and Policy and Master of Science in Computer Science and Engineering

Abstract

Time-of-day (or dynamic time-of-use, dToU) pricing is a mechanism by which system operators try to lower stress on the grid in times of high demand. The price for high demand periods is pre-set but the times of day they are applied is dynamic. Data on how residential consumers respond to the pricing scheme can inform more accurate models of consumption to maintain the integrity of the grid while lowering consumers' utility bills and optimizing renewable use. In this thesis, I analyze the data from a time-of-day pricing trial in London to see whether the treatment was effective in lowering consumption. I do this analysis using four different models and compare the accuracy of each and the results; an aggregated linear regression model, a multi linear regression model, an aggregated multi linear regression model, and a random forest regression time series model. I found that the time-of-day pricing during the trial was effective in lowering consumption and costs. A dependence on households' socio-economic status was observed.

Thesis Supervisor: Munther A. Dahleh
Title: Director, Institute for Data, Systems, and Society
William A. Coolidge Professor, Electrical Engineering and Computer Science

Acknowledgments

To my mother, forever my home and reason for all the successes I'll ever reach, you sacrificed so much so I could have more opportunities. I love you. To my grandmother, whose strength and resilience will always be with me. To my dad and to my uncles, whose guidance has gotten me through degrees and across borders. To the countless friends that make life more valuable, Shawndeez, Michaela, Mona, Yesenia, Danielle, and Joe your love and support mean a lot.

To the incredible Dean Blanche Staton, David Elwell, Dean Baluch, Ike Brochu, Dalton Jones, professor Dahleh, and Mardavij Roozbehani without whose support this thesis would not be possible. To the countless valuable friendships I've made in TPP and the support that got me through two pandemic graduate school years, Joonhee, Rebecca, Aleja, Disha, Hannah, Jack, Helen, Maya, Kevin, Taylor, Manon, your encouragements along the way got me through.

Contents

List of Figures

Chapter 1

Introduction

Electricity is unique in that its storage is prohibitively costly; this makes it essential for supply to, at least, meet demand at every second of every day[1]. Electricity generation, therefore, depends on a projected demand for every hour. In times of crisis where demand exceeds its projected amount, system operators — who are tasked with ensuring the reliable delivery of electricity to consumers, businesses, and industry — need to fall back on different methods to either lower demand or suddenly increase supply to meet that unexpected peak in demand. Increasing supply is quite costly as it would require buying power from other utilities in an open market. During periods of high demand — during a snow storm, a heat wave, or when renewable energy supply is low — system operators curb demand by incentivizing customers to lower their consumption. This tool is called demand response (DR). Utilities often think of DR as a virtual power plant since it operates by reducing load [28]. If at any point, demand exceeds supply, there will be a power outage. Power grid

[1]Supply can exceed demand, but at minimum it needs to be equal to it. In reality, there are always margins of error in place (usually 20%) as demand exceeding supply will lead to a power outage which is incredibly costly.

failures can be catastrophic and costly as events in February of 2021 in Texas showed. Department of Energy estimates that outages cost the U.S. economy about $150 billion annually [34]. Evaluating various interventions and incentive structures is critical in mitigating losses incurred in such scenarios.

1.1 Background and Motivation

A high-level motivation for implementing DR is that it can be shown that maximum social welfare is achieved when the retail price equals the marginal cost of generating and delivering energy [19, 20]. The barrier to demand response is technical and regulatory in nature. MIT's own Paul Joskow and Jean Tirole have shown that consumers' inability to access market transaction information in real time hinders retail competition[2] [17]. The prevalence of static pricing is two-fold. First, static pricing creates cross-subsidies. Consumers that consume most in high price hours are being subsidized by consumers that consume most in low price hours. For this reason, static pricing encourages excessive consumption during peak hours when cost of generation is higher than the retail static price [12]. Second, demand response increases retail competition which is to the consumers benefit. Hence, there's sufficient economic rationale for implementing DR. Quantifying the benefits of DR through studying data from trials can help justify the cost of installing smart meters and sway regulation towards dynamic pricing.

The challenge with evaluating the effectiveness of incentive structures that can be called upon to lower the strain on the grid in times of high demand is that it is impossible to know what customers would have consumed in the absence of said incentives, otherwise known as their counterfactual consumption or the customer

[2]Demand response are introduced in wholesale markets, not at the retail level.

baseline load (CBL). The difference between actual load and the CBL is considered the amount of load reduction, or the DR performance achieved by the customer [25]. Data collected from trials and pilots are used to help system operators find robust and accurate methods of estimating CBL. An accurate estimation of the counterfactual consumption is vital in deeming a DR incentive structure effective. High incentives put a heavy financial burden on the system operators; low incentives leave the customers dissatisfied and disengaged. Evaluating such incentive structures hence always includes counterfactual analysis.

In this thesis, I look at time-of-day pricing (also known as dynamic time-of-use), on of such incentive structures. Time-of-day pricing is a pricing mechanism under which the price per kilowatt-hour (kWh) depends on the time-of-day. The price bands are pre-set but the hour of the day they are applied to is dynamic. This is different from time-of-use (ToU) pricing, which targets predictable high demand periods in the week but is otherwise static, and critical peak pricing (CPP), which is not static but infrequent and targets only the highest demand periods of the year.

This pricing scheme reduces system stress during peak hours by increasing the price during those hours and lowering other times. This encourages consumers to reduce consumption during peak hours and shift some consumption to lower price hours. The research question raised here is hence, **whether this intervention can successfully help lower consumption during peak hours**. To answer this question, I analyzed data from a trial in London in 2013 to determine whether the implemented dynamic pricing scheme provided sufficient incentive for consumers to change their consumption habits. More details on the trial and the data set can be found below, in chapter 3. As expected, the time-of-day pricing was successful in lowering consumption during the high price hours. One of the features present in the data set is the socio-economic status of the households. I also examine significant

differences in the response among three socio-economic groups: affluent, comfortable, adversity, and conclude that the adversity group is least price sensitive and the comfortable group is most price sensitive.

In summary, the two questions this thesis works to answer are the following: 1. Did dToU lower consumption during the high hours? 2. How does socio-economic status affect response to dToU?

1.2 Overview of Related Work

Demand response literature approaches baseline estimation in one of four general ways: an economics approach in which consumers are maximizing a utility function or consuming based on a demand model [30], a machine learning approach that learns the exact patterns of usage for the particular consumer for which there is pre-intervention data [16, 25, 29, 35], a day matching approach which takes a historical look into the past and average consumption for non-event days with similar profiles; same day of week, temperature profile, etc. for an estimate of the baseline, or a regression approach which is often used for load prediction [31]. The latter two are the models that are most used in the industry [14]. Much of the current literature explores methods at the intersections of the above four general categories.

Additionally, the models are often either created for the purpose of baseline estimation or load forecasting. The baseline estimation methods help find the baseline from which treatment effects can be calculated. Load forecasting models are more often used by system operators to predict the amount of load on the grid. The two types of study, however, are very similar from a technical standpoint in that they attempt to find the underlying pattern of consumption and how that depends on other inputs to find consumption at a time for which ground truth data does not exist.

The sole difference is that baseline estimation is backwards looking and often in the context of historical data and load projection is forward looking. My thesis brings together the regression, matching, and machine learning approaches. It additionally draws from the work of Abadie et al. and the synthetic control model [5, 6], as well as Agarwal et al. for matrix imputation and completion [7]. Below, I dive deeper into some of the work that has informed my thesis.

Starting out with the economics perspective, Schneider et al. use an online learning model to find customer baselines [30]. For this purpose, they consider two customer demand models and investigate two natural objective functions for demand response programs for system operators. They conclude that the optimal demand response baseline is not necessarily the customer's counterfactual consumption. Sometimes, system operators benefit from a baseline that is not optimized per customer; the objectives for the are in conflict. A subset of similar economics papers derive energy demand models and how they are modulated by incentives, from first principles. Subbiah et al. found that daily consumption has an active portion and a passive portion [32]. They then use different data sets to validate the results obtained by this model against real-world data. The results show that the modeling framework is robust.

Demand response aims to influence the behavior of a subset of houses in aggregate. Synthetic control is a method that creates a 'synthetic' control group in cases where there is no explicit control group. Synthetic control is a particularly useful method in cases where we need to look at the aggregate affect on the outcome of a trial, especially where traditional regression methods are not appropriate — where the effects of a policy change is of interest, for example [5]. Abadie et al., in their 2003 synthetic control paper, constructed a 'synthetic' basque country without terrorism using data from a combination of neighboring areas to then be able to study

the effects of terrorism on the economy and the following cease fire [6]. Similarly, in a 2010 paper they estimate the effect of Proposition 99, a large-scale tobacco control program, on California's tobacco consumption [5]. The above qualities make synthetic control a very effective tool for finding the treatment effect of different pricing schemes.

Machine learning models work best if the data used to build the model — baseline estimation or load predition models — have other features that can help with the accuracy e.g. temperature, size of household, square footage, etc. These models can be fine-tuned for other data sets but lack explainability and transparency. A masters thesis from Montclair State University [16] explores the same data set along with weather data from the Dark Sky API. They use a Long Short-Term Memory model to predict consumption. Their main goal is to find the hyperparameters that lead to the most accurate consumption prediction using weather data as features; They don't include socio-economic status or electricity price as features in their model. Another machine learning model that is used for generating long-term forecasts based on pre-intervention data, for example, is the Temporal Fusion Transformer (TFT) model. The TFT model is an attention-based Deep Neural Network (DNN), optimized for great performance and interpretability. It has been benchmarked against traditional statistical models (ARIMA[3]) as well as DNN-based models such as DeepAR, MQRNN[4] and Deep Space-State Models (DSSM) and outperforms them all [22]. The original paper discusses the model in the context of a few experiments, predicting electricity consumption being one of the examples they use. They use the model on the UCI Electricity Load Diagrams data set, containing the electricity consumption of 370 customers – aggregated on an hourly level as in [35].

[3]Autoregressive Integrated Moving Average
[4]Multi-Quantile Recurrent Neural Network

In accordance with [29], they use the past week (i.e. 168 hours) to forecast over the next 24 hours. Regression models can take many forms, and are similar to machine learning models in that they can be used for load prediction but are relatively more transparent and less accurate.

Park et al. draw from the matching framework and also use machine learning for a data-driven approach to baseline estimation. They run the time series data through a self-organizing map (SOM) and then use k-means clustering to turn the data into K disjoint clusters. A SOM is an unsupervised machine learning algorithm that transforms a high dimensional space into a topology-preserving output space [21]. The output of the SOM is of much lower dimension compared to the original data set. They then split that data into K clusters. This creates clusters of houses that consume with similar patterns which allows for houses in the future to be mapped to the cluster they belong to. Lastly, they validate their model using data from Korea from 2012 to 2014 [25].

Several pieces of related work provide context for this research study. The most important is a dissertation from Imperial College London which is responsible for creating the data set for the Low Carbon London Project and initial analysis [31]. The author, James Schofield, explains experimental design principles (such as timing of "high" price events, potential flaws in the data set (such as drop outs), general motivation for the project (to allow for greater penetration of wind energy for the London Power grid), and the statistical methods and analysis used to examine the data after collection. Schofield sought to investigate the efficacy of dynamic pricing as a method for reducing the burden of high consumption hours on the power grid and in general, found that dynamic pricing is an effective tool to reduce electricity consumption. With his methods, he estimated that demand reduction during peak hours is within the range of 6.8-10.2%. The dissertation also considers socio-economic

groupings' potential to respond differently to dynamic pricing but did not mention specific p-values regarding consumption differences for each socio-economic group.

This thesis fits into a context of many other DR studies. Specific load shifting studies, such as one conducted over the summers in Texas, aims to understand how consumption patterns change as a result of DR pricing methods by examining appliance usage statistics [10]. In this experiment researchers estimated the Average Treatment Effect by comparing the control and treatment groups energy usage per appliance. In addition, this paper looks at responses to both pricing incentives and information incentives — for example, a text that asks households to consume less. They conclude that while pricing may improve economic efficiency, there's little evidence of response to the information treatments [10]. In another study, Ito finds that consumers do not necessarily respond to changing tariff schemes, which justifies further scholarship in this field to better understand what energy customers will best respond to [15]. Additionally, it is important to consider the ethics and equity of these tariff schemes, as examined in another study focused on the benefits of two-part tariffs [9].

The province of Ontario in Canada is the only region, apart from the country of Italy, to have rolled out smart meters to all its residential consumers and to be deploying time-of-use (ToU) rates to all customers that remain with the option[5]. The data from Ontario is very valuable in the context of studying ToU pricing and its effects on consumption trends. The Brattle Group prepared an analysis of Ontario's time-of-use rates for the Independent Electric System Operator. The goal of the study was to quantify the change in consumption, estimate the peak period impacts, and to estimate the elasticity of substitution between the pricing periods and the overall price elasticity of demand [8]. In their analysis, they estimated a model of

[5]Customers have the option to opt-out.

consumer behavior using the Addilog Demand System model to find inter-period elasticities of substitution. They also estimated a monthly consumption model to find an overall price elasticity of demand. The difference between this work and my thesis is that 1. they estimate an advanced economic model and 2. they estimated a general monthly consumption model. They use principle component analysis to reduce the dimensionality of the data which is also something I do.

1.3 Overview of Thesis

The rest of this thesis is organized as follows: chapter 2 goes through a historic overview of policies surrounding DR in the US, as well as defining different demand-side incentive schemes and estimation models. Chapter 3 outlines the specific trial and data set I use to quantify the effect of time-of-day pricing treatment, as well as limitations with the data set. Chapter 4 mathematically frames this question and outlines multiple regression models for finding the treatment effect, presents results for each, and quantifies their accuracy. Chapter 5 outlines a review of time series models and the details of a random forest regression model that learns the temporal trends as well as dependence on temperature. Chapter 6 outlines the conclusion, policy implications, limitations, and future work. Specifically, I will be comparing the results and accuracy of all the models; all show that the treatment was effective in lowering consumption during the high hours. This is impactful as it demonstrates data driven models of baseline estimation which is useful for system operators.

Chapter 2

Demand Response in the US: The Evolution of Forecasting Methods and Pricing Models

In this chapter, I outline ways demand response has evolved in the last twenty years. The first part of the chapter is focused on different forecasting methods and pricing models. The second half, is focused on policies around demand response and its increasing importance in the energy space. Lastly, I look at recent events in California and Texas to see whether the methods or policies have been effective.

2.1 Demand Response

Demand response is a method by which system operators are able to lower peaks in demand by offering time-based rates or other forms of financial incentives[1]. Con-

[1]Based on a definition from the Office of Electricity.

sumers are offered a financial incentive if they lower their consumption below what it otherwise would have been. In other words, whether or not they are awarded and by how much depends on a counterfactual consumption amount: it's impossible to know how much they would have consumed if the demand response incentive had not been offered. The very foundation of demand response, as a result, depends on an accurate forecast of what the consumption would have been in the absence of the incentive.

The main methods used in demand responses are real-time pricing, critical peak pricing, hedging consumption, or drawing from microgrids. I will expand on the way each operates below, as well as the ways in which each falls short. Which method is most optimal is a topic of debate among economists and policymakers. What introduces further complication is the degree to which the power market is regulated. Deregulated power markets operate much like commodity markets. Texas, for example, has an unregulated electricity market. For ease of analysis, I'll be eliminating the degree to which power markets are regulated.

2.1.1 Forecasting Models

A big challenge of demand response is forecasting the baseline from which reductions are measured. Almost always, historical data is used to find an estimate of the baseline. Forecasting algorithms typically look at consumption in the past to predict consumption in the future. There is the underlying assumption that commercial or industrial consumption is a function of the consumers and their behavioral patterns, hence future behavior is best predicted using consumption data in similar circumstances in the past.

There are many ways to manipulate historical data to generate a prediction of

consumption in the future. Generally, they follow the following three basic models:

- A prediction that is the average of the consumption over the past N days.
- A prediction that is based on the same days in the previous year.
- A prediction that is based on days with a similar profile (temperature, what day of the week, etc.) in the past.

These predictions are usually done in aggregate and not on a per-house basis. In addition, whether a consumer responds to demand response incentives, by how much they lower their consumption, and how quickly they respond to said incentive is studied and factored into how operators mediate peak hours. Again, given the aggregate and probabilistic nature of the analysis, it can't be depended on for certain. "The grid operator wants to have resources that it knows with near certainty it can call on to increase supply or reduce demand, " (Borenstein, 346)[2].

The above methods are not perfect. For instance, the first hot day of the season will inevitably have a higher consumption compared to the previous N days or even potentially compared to the consumption during the same time frame in a previous year. This makes it so that consumers that have helped keep the grid stable by lowering their consumption won't be rewarded because their forecasted amount was too low — they have lowered demand but not lower than the projected baseline. Regardless of the exact method, an estimated baseline can fail in a few ways:

1. If firms or customers know that their baseline is set based on previous peak days, they could strategically increase their baseline by consuming more during previous high-demand days as evidenced by a pilot program in Anaheim, CA

[2]Griffin, James M., and Steven L. Puller. Electricity Deregulation: Choices and Challenges. University of Chicago Press, 2009.

(Wolak, 2006). This would undermine the very reason behind the incentive. An inflated baseline would ensure that they are compensated more later.

2. Similarly, it might incentivize customers to consume more during baseline setting times to then be paid to lower their consumption. There was an instance where Camden Yards stadium turned on their lights on a day when there was no game. PJM (Pennsylvania, New Jersey, Maryland Interconnection) declared an emergency event immediately after. They were later paid to lower their consumption i.e. turn off lights that they did not need to have on in the first place.

3. Folks are disincentivized from investing in energy-efficient technology. This is because their overall consumption would be lower, but so would their baseline. In other words, the incentives reward the biggest delta between their 'baseline' and consumption during peak hours which does not necessarily lead to the most efficient results.

Forecasting is a difficult task because demand response is offered in times of crisis; winter storms, heatwaves, etc. for which there isn't realistic, dependable historic data. In addition, lack of certainty around the amount of reductions as a result of demand response measures and the timeline with which they will happen makes forecasting an even more important part of demand response. "[The grid operators] are less attracted to the idea of balancing supply and demand through price adjustment that yield small-quantity changes from thousands of customers, because none of those customers precommits to make a specific change under specific conditions. Rather, the responses to price changes are probabilistic, and the reliability of the aggregate response is due only to the law of large numbers applied to many independent buyers. To be concrete, if demand is exceeding supply and adjustment is

supposed to occur through a price mechanism, a grid operator has no one to call to assure that demand response occurs," (Borenstein, 346)[3].

Forecasting models are also heavily depended on to adjust electricity generation. Recall that the integrity of the grid is reliant on supply always exceeding demand[4]. Whatever is demanded but not generated would need to be bought from peaker power plants[5] at the spot price. The spot price of electricity is the price at which electricity can be bought or sold for immediate delivery. The senior author of Spot Pricing of Electricity and late MIT professor, Fred C. Schweppe, created the concept of spot pricing and proved, again, that "the forecast is always wrong!" (Schweppe, v).

2.1.2 Real-Time Pricing (RTP)

Real-time pricing, also known as dynamic pricing, is the situation in which the hourly price of electricity depends on the utilities' production cost in that time frame. This price is surely higher in times of high demand as peaking power plants are more expensive to run than base load plants[6]. A dynamic pricing scheme is used both as a result of an unregulated market or as a demand response method.

As mentioned above, dynamic pricing is a way to lower the effects of issues raised from an inaccurate baseline estimate. This solution has its problems: dynamic pric-

[3]Griffin, James M., and Steven L. Puller. Electricity Deregulation: Choices and Challenges. University of Chicago Press, 2009.

[4]Within a safe margin, often 20%.

[5]Peaking power plants or peaker plants are power plants that are spun up during peak hours. They typically use natural gas, are less efficient, and often have higher emissions per kilowatt-hour power generated. This is due to the fact that they are only used occasionally. Hydropower is also a common peaking power source. Water is pumped to higher altitudes in times of low demand using extra generated power for a peaker event in the future.

[6]Base load plants operate continuously at near full capacity. They are cheaper as they use low-cost fuels. These plants supply the majority of the expected demand in the network; their outputs are quite inelastic.

ing makes it so that electricity is treated like all other commodities with its price increasing in times of high demand. The potential for extremely high prices makes this solution impractical as it exposes the consumers to too much risk. This is what left folks that were 'lucky' enough to have electricity in Texas in winter 2021 with extremely high electric bills.

2.1.3 Critical Peak Pricing (CPP)

Critical Peak Pricing (CPP) programs give customers low prices throughout most of the year but have quite high prices in critical times — when demand is unexpectedly high or supply is low; usually, these cases happen at the same time. The higher price in the peak hours incentivizes consumers to lower their consumption: either shift to off-peak hours, fully reduce consumption, or rely on a backup generator to meet needs. Studies show that CPP programs yield substantial demand reduction and satisfaction among customers both in residential and commercial settings. Similar to RTP, CPP may leave consumers with extremely high electric bills. Additionally, CPP programs are more dependent on an accurate baseline estimate which for the reasons outlined above is difficult to capture.

2.1.4 Hedging

CPP and RTP or variations of the two are pricing schemes that are most often used in demand response. I outlined the shortfalls of each above: they both can lead to extremely high electric bills in peak hours. A solution to manage the amount of risk customers are exposed to is hedging, where consumers themselves estimate their future consumption and purchase a certain number of kilowatt-hours of power in advance — before other metrics, such as the weather, are known. The energy that

would be purchased in advance is lower cost, making hedging a solution that can control the volatility of a customer's electric bill.

Hedging still incentivizes customers to lower their consumption in peak hours because any kilowatt-hour above their previously purchased amount would have to be purchased at the high spot price. Additionally, any kilowatt-hour that they consume less than the hedge quantity, they can sell back at the spot price. Customers hedging their own consumption is a good start but still inadequate. It's unrealistic to expect individuals to accurately predict their consumption, especially in times of crisis, when system operators' forecasting models, with all the data that they have available to them, fail. Hedging could still leave consumers with the need to buy megawatts at the spot price which would leave them with a high electric bill.

The issue is that hedging is positively correlated with prices; when prices are high, it means that it's a peak event, therefore, demand is high and the hedging quantity should have also been high because the grid is likely stressed. A solution to that is to over-hedge. A better solution is a hedge quantity that fluctuates proportionally with the overall system demand; e.g. on days that the system demand is higher, the hedge quantity is also proportionally higher. This lowers the amount of electricity that might be needed at the spot price.

Most of these solutions, hedging included, allow consumer choice. Larger industrial consumers can use more sophisticated consumption or hedging strategies to lower their electric bills. Smaller consumers can pick a default plan from their retailer.

It's important to note that dynamic pricing with hedging is superior to programs that pay consumers to lower their demand relative to an estimated baseline. As outlined in the forecasting model section, this estimated baseline is a function of that customer's consumption in the past and, therefore, at the risk of baseline manipu-

lation, adverse baseline selection (participants can choose a program and a baseline algorithm that works best for them), and participation selection (consumers for whom none of the programs are personally beneficial can opt-out of all demand response programs and continue purchasing electricity at the constant flat rate). What this means is that dynamic pricing combined with hedging prevents overcompensating as a result of baseline manipulation, as well as lowers risks for both consumers and retail providers.

2.1.5 Microgrids

Some residential and commercial buildings now have their own mini-grids; some have solar panels that, unlike power plants, can save electricity, some even have their own backup generators. Microgrids are another solution that can come to the rescue in times of peak demand. A microgrid, as the name suggests, is a small, local grid that can disconnect from the main grid and operate autonomously in times of crisis. Microgrids have power generation and storage capabilities and can power a building, campus[7], or even a small town[8]. Recent research attempts to bring in the power stored in electric cars as a source in times of high demand.

2.2 Background on the US Power Grid

Getting electricity from the power plants to our homes is done in three main steps: generation, transmission, and distribution. A combination of the transmission and

[7]New York University has generated power on site since the 1960s. NYU's power plant was successful in powering the campus during Hurricane Sandy which kept the campus lit, unlike most of the New York downtown area.

[8]The microgrid in Fort Collins, Colorado is part of a project (Fort Collins Zero Energy District, FortZED) where the district plans to create as much power as it consumes.

distribution steps is referred to in North America as the *power grid* or the *grid*.

The US Grid dates back to 1882 — when Thomas Edison launched the first power plant on Pearl Street in New York City. The grid has expanded quite a bit in size since then but the underlying structure has remained unchanged.

With 7700 power plants, 3300 utilities, and over 2.7 million miles of power lines, the grid has been called the largest machine in the world. The US grid, however, has three main operational components: the Eastern, Western, and Texas interconnections. The grid is considered a 'natural monopoly' given its complexity and costly infrastructure; therefore, historically, the utilities controlled every step of the process. Regulated markets still function the same way, with power utilities owning the entire electricity supply chain from generation to distribution.

In 1978, Congress passed the Public Utility Regulatory Policy Act (PURPA) which required utilities to buy power from independent third parties that could generate power for cheaper. 1992 further deregulated the power market; the 1992 Energy Policy Act separated power generation (wholesale market) from transmission and distribution — utilities kept their monopoly over local electricity distribution. This made it so that generators would now sell power to utilities and retail electricity suppliers on the wholesale market (business-to-business) at rates set in a competitive bidding process who then sold the power to consumers on the retail market (business-to-consumer) with the cost of distribution and transmission from distribution utilities added.

2.3 The 2005 Energy Policy Act and Order 745

The 2005 Energy Policy Act (EPAct 2005) transformed the power sector. Following the 2000-2001 power outages in California, rising energy prices, and dependence on

foreign oil, the 2005 Energy Policy Act attempted to address increasing concerns about energy security, environmental quality, and economic growth (Congressional Research Services). Passed by Congress in **July 2005** and signed into law by President George W. Bush in **August 2005**, EPAct designated the Department of Energy's Federal Energy Regulatory Commission (FERC) as the primary authority over power generation and transmission across the US.

Based on the Energy Policy Act of 2005, the Secretary of Energy shall be responsible for "(1) educating consumers on the availability, advantages, and benefits of advanced metering and communications technologies, including the funding of demonstration or pilot projects; (2) working with States, utilities, other energy providers and advanced metering and communications experts to identify and address barriers to the adoption of demand response programs; and (3) not later than 180 days after the date of enactment of the Energy Policy Act of 2005, providing Congress with a report that identifies and quantifies the national benefits of demand response and makes a recommendation on achieving specific levels of such benefits by January 1, 2007."

Section D of EPAct 2005 includes detailed responsibilities for FERC to reform transmission rates. In summary, the commission is responsible for the following:

- Establishing incentive and performance-based rate treatments in the year following the enactment of this section.

- The above is to assure reliable and low-cost transmission of power by reducing congestion. It shall also promote the economically efficient generation of power "by promoting capital investment in the enlargement, improvement, maintenance, and operation of all facilities for the transmission of electric energy in interstate commerce, regardless of the ownership of the facilities."

- "In the rule issued under this section, the Commission shall, to the extent within its jurisdiction, provide for incentives to each transmitting utility or electric utility that joins a Transmission Organization. The Commission shall ensure that any costs recoverable pursuant to this subsection may be recovered by such utility through the transmission rates charged by such utility or through the transmission rates charged by the Transmission Organization that provides transmission service to such utility."

- All rates must be just, reasonable, and not unduly discriminatory or preferential.

EPAct 2005 reaffirmed a commitment to competition in wholesale power markets as national policy and put a lot of emphasis on demand response and clean energy. A number of orders followed; on **July 20, 2006**, Order 679 attempted to address the responsibilities outlined in Section D — to bolster investment in the nation's transmission infrastructure which would benefit consumers "by ensuring reliability and reducing the cost of delivered power by reducing transmission congestion," (Order 679).

In **August 2006** FERC issued the Assessment of Demand Response and Advanced Metering as required by EPAct 2005. In the report to congress, FERC outlines the regulatory barriers to "improved customer participation in demand response, peak reduction, and critical peak pricing programs." Some of the most significant regulatory barriers are:

1. The disconnect between retail and wholesale rates

2. Demand response disincentives for utilities

3. Lack of incentives for utilities to use "enabling technologies"

4. The need for additional research on the cost-effectiveness of pricing models

5. State-level constraints to offering greater demand response

6. Barriers for third-parties to offer demand response and inaccessible demand response data for third-parties

7. The need for better federal-state coordination on demand response offerings

The report concludes that "demand response needs serious attention." Their suggested remedies for the state of demand response are 1) to explore ways to increase presence in the wholesale market; 2) better coordinate state commissions with utilities; and 3) to remove regulatory barriers that hinder participation in demand response, peak reduction, and critical peak pricing programs.

On **June 22, 2007**, FERC issued an Advance Notice of Proposed Rulemaking. Among many other things, it sought public comments on "the role of demand response in organized markets, including greater reliance on market prices to elicit demand reductions during power shortages."

Order 719 issued on **Oct 17, 2008**, finalized regulation that leveraged demand response to improve the competitiveness of organized wholesale markets. EPAct 2005 had tasked FERC with benefiting consumers by boosting competition in organized wholesale markets and ensuring just and reasonable prices. On **February 22, 2008**, FERC issued another Notice of Proposed Rulemaking; this time they were addressing "demand response and market pricing during a period of operating reserve shortage" among other things. **July 16, 2009**, saw Order 719-A which affirmed in part and granted in part rehearing of Order 719. Order 719-B, issued on **December 17, 2009**, denied rehearing Order 719-A and finalized its determinations of pricing mechanism during reserve shortage in organized markets. The reforms in Order

719 attempted to treat demand response as other resources. It instructed Regional Transmission Operators (RTOs)[9] to allow demand response into wholesale markets. This allowed entry of a negawatt[10] of energy — a saved megawatt — into wholesale market auctions.

On **March 15, 2011**, FERC issued Order 745 that took Order 719 a step further. Order 745 instructed RTOs to compensate a negawatt of energy at the same rate as a megawatt of energy. This was foundational as it was the first time FERC had determined a price mechanism for demand response and not only that now demand response was part of the wholesale market (established in Order 719) but that it was worth the same as a megawatt of power.

Order 745 proved controversial. Energy Economists such as William Hogan[11], Raymond Plank Research Professor of Global Energy Policy and research director of the Harvard Electricity Policy Group (HEPG), openly criticized the choice to price negawatts and megawatts of electricity at the same rate — at the locational marginal price (LMP)[12]. The criticism stems from the fact that negawatts of power should

[9]Wholesale markets are managed by nonprofit regional transmission organizations (RTOs), which ensure that the grid remains reliable and that wholesale power prices remain "just and reasonable" through the use of competitive auctions. Nine exist in North America, covering about 60 percent of the US power supply.

[10]Amory Lovins, an American physicist and a big promotor of energy efficiency and the use of renewable energy sources, noticed a misprint in a report of the Colorado Public Utilities Commission in 1989: negawatt for megawatt (MW). He borrowed the term to describe a unit of power saved — a negative megawatt — through conservation or increased efficiency. (from "Negawatt Hour." The Economist, The Economist Newspaper, www.economist.com/business/2014/03/01/negawatt-hour.)

[11]Hogan has written extensively about Order 745 and more generally about how demand response should be priced. Some of those papers are as follows: Providing Incentives for Efficient Demand Response (October 2009), Demand Response Pricing in Organized Wholesale Markets (May 2010), Implications for Consumers of the NOPR's Proposal to Pay the LMP for All Demand Response (May 2010), Demand Response Compensation, Net Benefits, and Cost Allocation: Comments (November 2010), Demand Response: Getting the Prices Right (February 2016).

[12]Locational marginal price recognizes that the cost of making a unit of electricity available for purchase can vary greatly by location (William Hogan, Electric Transmission: A New Model for Old Principles, The Electricity Journal, vol. 6, no. 2, p.18, 1993).

not be valued at the same rate as megawatts of power. Demand response is often cheaper than even the cheapest of power generations. Order 745 makes it so that someone curbing their demand not only saves how much they would have spent on that megawatt of power but earns the same amount for generating negawatts priced at the same rate. Richard J. Pierce Jr., a professor at George Washington University Law School [18] puts the failure in Order 745 very simply and eloquently: "[there is] an explicit 'reward' for conservation in addition to the market-based 'reward' the consumer gets as a result of a decision to decline to purchase a unit of electricity," (Pierce Jr., 2011).

This inconsistency alarmed power generators. The Electric Power Supply Association (EPSA), the coalition of major power generation owners, sued FERC claiming that the commission had overstepped its jurisdiction by meddling with retail markets, the states' domain. The American Public Power Association (APPA), the National Rural Electric Cooperative Association (NRECA), and the Edison Electric Institute (EEI) joined EPSA in the lawsuit.

In **May 2014**, the U.S. Court of Appeals for the D.C. Circuit sided with EPSA and ruled that demand response was a retail-side transaction, hence under states' utility commissions' jurisdiction. The court further stated that even if FERC had jurisdiction to issue Order 745 under the Federal Power Act, it would have been vacated as the choice to use LMP is "arbitrary and capricious" and FERC has failed to address the concerns of the order leading to unjust and unreasonable rates that overcompensate demand response resources.

The disagreements between either side continued. EPSA believed the Order was overcompensating demand-side resources and would hence result in premature closure of power plants. Supporters of Order 745 argued that demand response has always been valuable; it's always been more valuable to curb demand than to build

a peaking power plant that is used a few hours out of the year[13] [13].

The lawsuit was appealed to the Supreme Court. On January 25, 2016, the Supreme Court ruled to uphold Order 745. In a 6-2 decision, justices concluded that FERC was within its jurisdiction to be setting prices for demand response resources in the wholesale market. The Federal Power Act not only gives FERC jurisdiction over the wholesale market rates but also rules and practices 'affecting' the wholesale market. Justice Elena Kagan, who delivered the opinion, wrote "FERC has the authority — and, indeed, the duty — to ensure that rules or practices 'affecting' wholesale rates are just and reasonable."

Additionally, an important distinction is that Regional Transmission Operators (RTOs) [2] coordinate the transfer of power between states. They operate on a multi-state grid and are as a result outside of the state's jurisdiction and regulated by the FERC. On whether FERC was regulating retail markets, Kagan said the answer is unambiguously no: "When FERC sets a wholesale rate, when it changes wholesale market rules, when it allocates electricity as between wholesale purchasers — in short, when it takes virtually any action respecting wholesale transactions — it has some effect, in either the short or the long term, on retail rates. That is of no legal consequence."

With the supreme court ruling, demand response had a chance to grow as FERC intended even at the expense of the utilities overcompensating demand response resources and even if economically inefficient. It was forecasted[14] that ruling against Order 745 would cut the demand response industry growth in half and cost the US demand response market 4.4 billion dollars in revenue the following 10 years after the

[13]Said Audrey Zibelman, chair of the New York Public Service Commission.

[14]In a report by Greentech Media. I was unable to find the report as the company has since been sold and the link to the report redirects to Wood Mackenzie's home page but this article summarizes the report.

ruling. It's beyond argument that the Supreme Court ruling played a foundational role in the growing importance of demand response and efficient energy.

2.4 California August 2020, Texas February 2021

The blackouts in 2000-2001 in California were one of the reasons EPAct 2005 was signed into law. Looking at different demand response forecasting methods, their shortfalls, and the trajectory of demand response policies that led to their increased importance, one question remains: how effective are demand forecasting and demand response *today*?

The blackouts in California in August 2020 and in Texas in February 2021 have been humanitarian and economic disasters. Both events were a result of unexpected extreme weather, which is more common due to climate change.

The California Independent System Operator (CAISO) did a preliminary root cause analysis [11] on the rotating outages in August 2020. The report pointed to three reasons: 1) the extreme heat created circumstances that fell outside of what can be handled with existing resource planning 2) the transition to an efficient, green resource portfolio makes it difficult to bring in additional supply in the early evening hours[15], and 3) practices[16] in the day-ahead market exacerbated the supply challenges. On August 14, California experienced 1.4 to 2 gigawatts of natural gas fleet outages. This sudden outage wasn't able to be replaced; out of the 1.5 gigawatts of capacity available in demand response in California only 200 megawatts are available in less than 15 minutes. CAISO is itself looking to increase demand response

[15] California is relying more and more on solar power and closing natural gas and nuclear power plants.

[16] under-scheduling of demand in the day-ahead market by scheduling coordinators, convergence bidding masking the tight supply conditions, and the configuration of the residual unit commitment market process.

flexibility and reliability in preparation for summer 2021. This points to a need to develop demand response methods and technologies that are dependable, flexible, and responsive irrespective of the time or time frame of the crisis.

On August 14 and 15 2020, in addition to extreme heat, California experienced two natural gas fleet outages of 400 megawatts and 470 megawatts. This sudden outage wasn't able to be replaced; out of the 1.5 gigawatts of capacity available in demand response in California, only 200 megawatts are available in less than 15 minutes (a design choice[17]) and the load needed to be dropped in a 12-minute window. To better understand the rolling blackouts let's take a survey of the resources available. 2.2 gigawatts of behind the meter resources are available: 1.5 gigawatts of demand response load-shedding[18], 450 megawatts of behind the meter storage both residential and commercial and industrial (C&I), 100 megawatts of EV charging flexibility, and 160 megawatts of natural gas and diesel microgrids. On the one hand, flex alerts, the most traditional and least predictable form of demand response, were utilized and shed 4 gigawatts[19]! On the other hand, much newer demand response technology such as Sunrun and Tesla were also utilized. Other resources were brought in during the two days; 180 megawatts from data centers and microgrids from the navy. During the rolling blackouts, diesel backup generators were given emergency authorization. Diesel and natural gas generators added 950 megawatts of capacity. The emergency authorization points to cycles of climate change.

In short, the different demand response resources were utilized over the two days, but not all at once. They helped in aggregate and functioned as designed. A main takeaway is that the resources need to be further integrated. CAISO is itself looking

[17]They are designed around a 30-minute response time.

[18]Which isn't really automated, hence the 30-minute time frame.

[19]Flex alerts are a great example of a commons problem because folks should accept that momentary lowering of the AC in aggregate can help keep the grid secure.

to increase demand response flexibility and reliability in preparation for summer 2021. This points to a need to develop demand response methods and technologies that are dependable, flexible, and responsive irrespective of the time or time frame of the crisis.

Many factors are to blame for the events in February 2021 in Texas: Texas's semi-isolated grid, unregulated market, wholesale market design, and record colds that led to frozen wind turbines, but most importantly the failure of its natural gas system. Frozen natural gas combined with a peak in demand caused gas prices to skyrocket. This is an indication that managing peak demand in electricity is not independent of other industries. It's important to think of increasing reliability in the natural gas industry as a part of the future of demand response in Texas's isolated grid.

The debate surrounding solutions involving supply or demand continues. Severin Borenstein, the Faculty Director at the Energy Institute at Haas, writes about an increasing consensus among economists, grid operators, and utility managers that paying people to lower their consumption is not working. An alternative approach, such as dynamic pricing with hedging, is more impactful, cost-effective, and less risky. This demonstrates the need for refining the policies around demand response and adjusting the grid infrastructure in a way that aligns all stakeholder incentives rather than methods that overcompensate consumers based on inaccurate baselines and aren't dependable for system operators.

2.5 EVs' Impact on Demand Response

With the improvement of EV batteries, an increasing number of charging stations, and lower overall cost, it's expected that the number of EVs will increase. According to data from the Edison Electric Institute, there will be 18.7 million EVs on the road

by 2030 in the US. They are projected to make up 20% of new car sales in the US annually until then. This will change the demand response landscape. EV charging needs to be planned for as it can create a peak during post-work hours, for example. EV demand response pilot programs are currently in place. Some offer free chargers in exchange for reduced, delayed, or shifted charging during DR events. EV charging is ideal for time-of-use as customers may be willing to give up control of their chargers if they can have their cars charged overnight or during renewable peaks for a lower rate. A guarantee to have EVs charged by a certain time and allowing customers to opt-out if and when they want is important according to VP and Chief Innovation Officer of Green Mountain Power.

Most importantly, however, EV batteries can act as microgrids. EVs not only reduce emissions from burning fossil fuels but can also be thought of as storage and a backup option. Research from the University of Bergamo shows that using EVs as microgrids in peak times can help lower the market-clearing price and increase demand response performance.

Lastly, as EVs help decarbonize the transport sector, it's important to continue the push to decarbonize the electric sector. Plots comparing emissions from the two sectors show different trajectories in richer vs poorer regions. One possible explanation is that rich nations are exporting the emissions from their electric sectors to seem more locally sustainable but aren't able to do that for the transport sector for obvious reasons. It's important to develop technology that allows EVs to be charged by renewables and without the need for electricity generation through means that increases emissions (peaker plants during DR events, for example).

2.6 Conclusion

The importance of demand response was realized in the wake of the 2000-2001 California blackouts. The regulatory landscape around demand responses in power markets has gone through a major shift since then; it's clear that both policymakers and energy economists alike understand the importance of being able to navigate times of stress on the grid. Current demand forecasting and response methods have come a long way but still fall short to accurately and dependably lower demand in peak hours. Recent events in California and Texas show that current demand response measures, though helpful, fall short of providing a perfect solution. It's essential to learn from the ways in which current forecasting and demand response methods could have performed better and iterate on the policies around such topics to prevent disastrous blackouts in the future.

Chapter 3

Low Carbon London Smart Meter Trial

This chapter describes the specific details of the trial and data set used in this thesis, as well as the research question the trial helps us answer. I go through the trial design, data exploration and cleaning, and close with the limitations of the data set and ways to analyze the data given the existing limitations.

3.1 Trial Description and Design

Programs such as Low Carbon London (LCL) were created following the Climate Change Act of 2008 [26] and a commitment to lower carbon emissions to 20% of 1990 levels by 2050. The LCL program was funded in 2010 under the Low Carbon Networks Fund (LCNF) tier 2 scheme by the amount of £21.7 million with an additional £6.6 million of funding contributed by program partners [1]. Part of their goal was to understand demand side response; specifically, the mission was to gather data

on the performance of smart grid technologies. To this end, a dynamic time-of-use (dToU) trial in collaboration between Imperial College, UK Power Networks[1], EDF Energy[2], and some other stakeholders was designed, implemented, and analysed. I used the data set from this trial for my analysis.

Here, I will discuss the data set's contents, which were divided based on control and treatment, with categorizations by socio-economic status. The data set with which I conducted my analysis is accessible here. It was collected from the UK Power Networks between November 2011 and February 2014 for 5,567 households. A subset of 1,100 of the households, recruited as a "balanced sample representative of the Greater London population," were subject to dynamic time-of-day pricing — their electricity tariff would be one of three previously announced rates (high, normal, or low) depending on the time-of-day. The tariff prices were given a day ahead via the Smart Meter IHD (In-Home Display) or text message to their mobile phone. Customers were issued high (67.20 pence/kWh), low (3.99 pence/kWh) or normal (11.76 pence/kWh) price signals, and the times of day they applied. The prices were chosen such that a consumer's bill remain unchanged in the case they did not respond to DR events. The rest of the households which made up the control group, were priced on a flat rate tariff of 14.228 pence/kWh. The data set includes consumption readings that were taken every half-hour (kWh/hh), date and time, two socio-economic consumer classifications, whether the households are in the control or treatment group, and a unique household identifier.

The two socio-economic classifications are from CACI's Acorn classification; one is more granular and segments the households into 17 groups, the other is less granular and segments the households into three groups, adversity, comfortable, and affluent.

[1]The London distribution network operator (DNO) and the lead program partner
[2]Retail energy supplier

CACI is a UK based consulting firm that offers data and technology solutions to public and private clients. Per their website, Acorn is a consumer classification that splits the population into 62 different types and provides details on consumer characteristics. To better understand the socio-economic Acron groupings, I referred to the publicly available User-Guide [3]. However, the guide was not very technically detailed and did not explain the groupings present in the data set or the reason and meaning behind any such groupings. I reached out to the CACI, who created and continues to update Acorn groupings, and they kindly provided me with the most recent technical guide [4].

There are several key takeaways from these documents. First, the Acorn Groups were initially relied upon as a data collection measure to ensure that trial and testing groups were created to be representative of London as a whole as well as to stratify responses during the analysis explained in the Schofield thesis [31]. Additionally, the Acorn groupings used in the 2013 trial were generated specifically for this data set and therefore, are slightly different than the groupings in the User Guide and Technical Guide. The Acorn classifications used in the data set are not the same, freely available geo-demographic/zip code-based classifications described in the user guide. Schofield uses 87 Acorn types, which can be categorized in 17 Acorn groups, and distilled down to 5 Acorn categories. In the User-Guide there are 62 types, 18 groups, and 6 categories. The Low Carbon London (LCL) project appears to have commissioned a study specific Acorn classification system, which accounts for the difference.

The trial was double opt-in; all houses present in the trial opted into sharing their data. This was true both for the treatment dToU group and the non-time-of-use (non-ToU) group [31]. Additionally, the treatment group opted into the treatment group and being subject to time-of-day pricing for the calendar year of 2013. The houses

under study had to have smart meters installed, it is therefore assumed that the sample is skewed towards technology enthusiasts, those that were reachable through recruitment methods, and those who had the time to have the installation take place (required an engineer to visit home) [23, P.11]. The sample recruited is, to some degree, biased towards 'early adopters' of smart meters [23, P.11]. The treatment group was offered further incentives. Some of the incentives are outlined below [27, P.26]

- A guarantee that they will be reimbursed at the end of trial if they are worse off on the dToU tariff than they would have been on their previous tariff.
- £20 for returning the appliance survey.
- Assurances regarding how many hours would be charged at the high price band.
- £100 for signing up to the dToU tariff .
- Another £50 for staying on the dToU tariff until the end of trial.
- £20 for returning the consumer dToU tariff survey at the end of the trial.
- Entry into a prize draw after completion of the post trial survey.

Trial recruitment began in 2011 and continued through 2012. For this reason, the number of houses for which I have data is increasing in 2012 as the recruitment progressed. Figure 3-1 shows how the number of houses grew over the time period for which data exists. Over 2011 and 2012, the recruitment was actively taking place hence the growing number of houses for which we have data. The slight decline in 2013 is due to houses that withdrew from the trial. Although we have data from November 2011 to February 2014, the treatment group only went through the time-of-day pricing over the calendar year 2013.

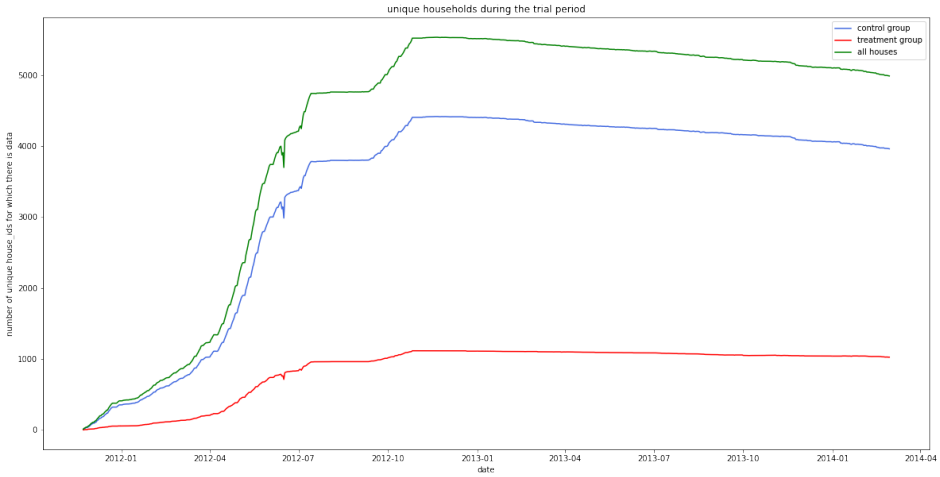

Figure 3-1: Number of unique houses for which data exists during the trial period.

The relative population of the two groups was decided to guarantee statistical significance. Additionally, recruitment was continued until both groups had a representative sample of London in terms of both location and Acorn groupings which signify socio-economic status. If a class of consumer was found to be underrepresented, recruitment was intensified within this group until the correct ratio was achieved [31].

3.1.1 Data Ingest, Pre-Processing, and Exploration

The data includes the following: a unique `house_id` per household, two Acorn groupings: one that is more granular and another that segments the population into three subgroups, whether the household was in the treatment or control group, half-hour level consumption data in kilowatt-hour (kWh), and the date and time of the consumption measurement. The data for November 2011 until February 2014 is stored in

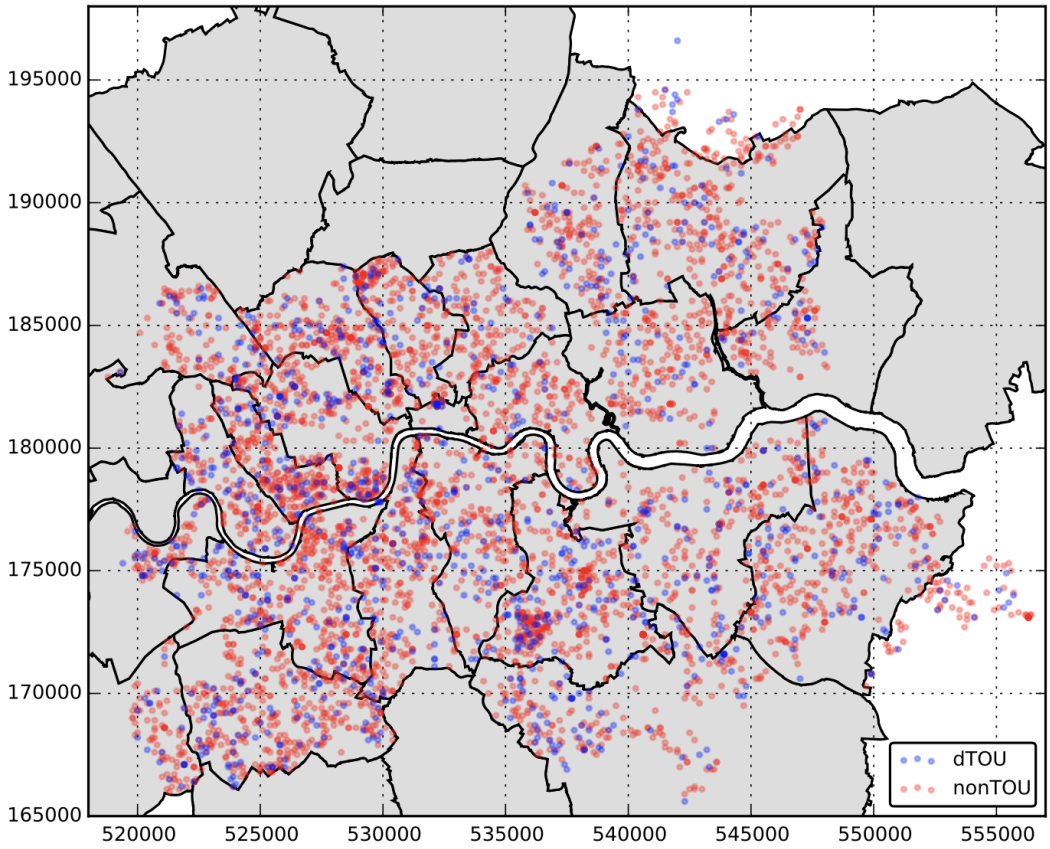

Figure 3-2: Trial household sample locations overlaid on the borough boundary map of Greater London. Map data from the Greater London Authority. This shows that the treatment and control group were representative samples of Greater London.

168 CSVs. Additionally, for the year the treatment is in effect (2013), there is a CSV mapper that demonstrates which hours in 2013 had low, normal, or high price points for the treatment group. In the pre-processing step, I combined all the 168 CSVs into two gzip files: one with all the consumption data (`house_id`, whether or not the house is in the treatment or control group, consumption in `kWh/hh`, and the date and time) and another with all the Acorn data (`house_id` and the two Acorn groupings per household). The `house_id` was shared among the two files which helped map

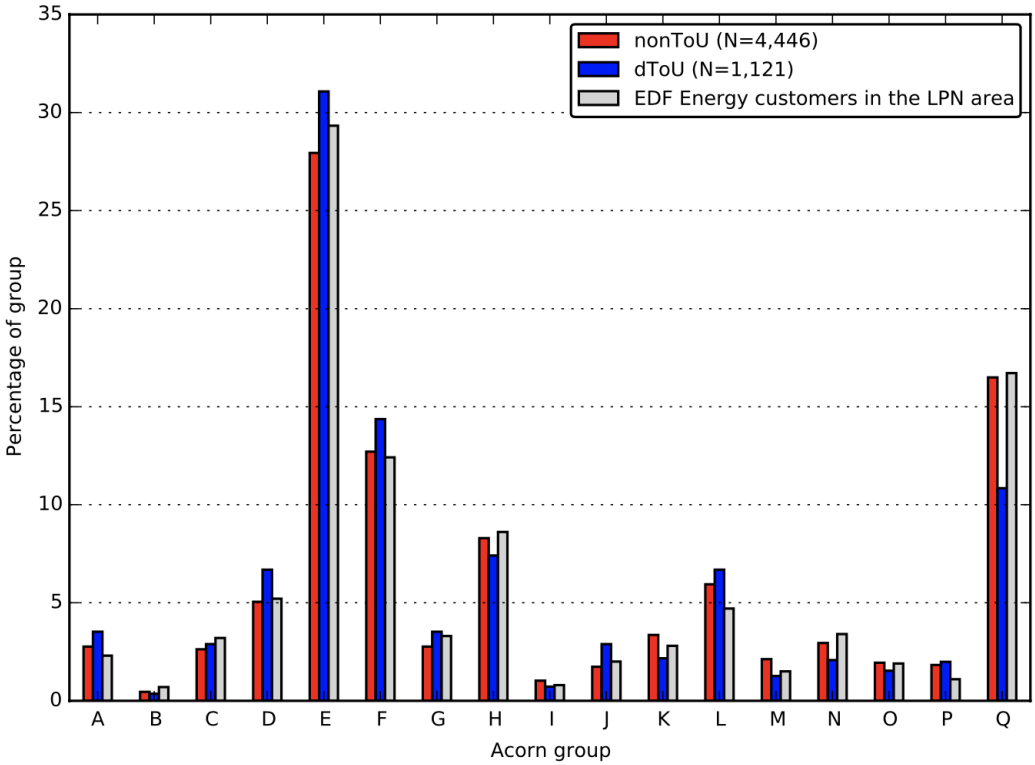

Figure 3-3: Proportions of Acorn groups for the dToU group the nonToU group and EDF Energy customers in the London Power Networks (LPN) area. The increasing alphabetical ordinals of group labels' loosely correspond to increasing household wealth e.g. {A, ... , Q}. This shows that the treatment and control group had similar distributions of customers in different socio-economic groups.

the data between the two tables. I then segmented the consumption data into the years for which I had data. This allowed me to only import the fraction of the data with which I was running an analysis on at the time.

Clustering Houses

An interesting line of inquiry is clustering houses based on their patterns of consumption. This is helpful in matching algorithms as in can help more closely find

a baseline from historic data for houses within that cluster. It can also be used by system operators for targeted DR. Clustering houses that respond similarly to a DR event and knowing whether and to what extent a cluster of houses respond to a DR event is useful information for system operators as they can reliably call upon that cluster of houses in the case of an event.

I used different clustering methods to cluster houses in this data set based on the control group's consumption pre-intervention: k-means clustering, PCA[3], TSNE[4]. I also tried clustering on the frequency responses which I found fast Fourier transform. PCA, TSNE, and Agglomerative clustering on the resulting data set was also inconclusive. This is reason for sticking to the pre-existing socio-economic clusters.

3.2 Hypotheses

Revising from chapter 1, the two questions in this thesis are whether dToU was effective in lowering consumption during the high price hours in this trial and whether socio-economic status affects level of response to the intervention. With these questions in mind, the following two hypotheses will be tested.

The first hypothesis is that the treatment was effective in lowering consumption in the high hours. Here, the ATE is the effect of the treatment on **consumption values**.

$$\begin{aligned} H_0 &: \text{ATE(high price hours)} = 0 \\ H_1 &: \text{ATE(high price hours)} < 0 \end{aligned} \tag{3.1}$$

This hypothesis is grounded in the fact that demand response mechanisms, dToU

[3]principal component analysis
[4]t-distributed stochastic neighbor embedding

included, are created for the soul purposes of lowering consumption during peak hours.

The second hypothesis is that the higher the socio-economic status, the lower the price sensitivity. Here, the ATE is the effect of the treatment on the **cost** of electricity.

$$\begin{aligned} H_0 &: \text{ATE(affluent)} = \text{ATE(comfortable)} = \text{ATE(adversity)} \\ H_1 &: \text{ATE(affluent)} > \text{ATE(comfortable)} > \text{ATE(adversity)} \end{aligned} \tag{3.2}$$

This hypothesis is grounded in the fact that the adversity subgroup, given their socio-economic status, is most likely to shift their consumption around the price signals. However, the ability to shift demand also depends on the possibility of doing so. It's possible that the adversity group contains households that due to work hours are simply unable to shift their consumption around the price signals. Another possibility is that affluent subgroups have smart appliances and are more sophisticated consumers. As a result, it's possible that lower socio-economic groups will not be able to take advantage of dToU or ToU pricing.

3.3 Treatment Effect

The goal in this thesis is to be able to use the in sample and out of sample data from the control and treatment groups to quantify how effective the time of use pricing was in lowering consumption in high price hours. In order to be able to quantify this effect, we need to estimate a counterfactual consumption for the treatment group for the treatment period. I will expand on the specific details of the different methods I used for finding the counterfactual consumption in chapter 4.

Generally speaking, treatment effect or causal effect of the treatment on the outcome for unit i is the difference between its two potential outcomes:

$$Y_{1i} - Y_{0i}$$

Where Y_{1i} is the potential outcome for unit i with treatment and Y_{0i} is the potential outcome for unit i without treatment. The fundamental problem of causal inference is that both potential outcomes (Y_{1i}, Y_{0i}) cannot be observed. A large amount of homogeneity would solve this problem. For example, if one could assume that (Y_{1i}, Y_{0i}) is constant across individuals or time. Randomly assigning people to the two groups would make that assumption possible and would force the bias term to zero.

As expanded in section 1.2, synthetic control is a method that creates a 'synthetic' control group in cases where there is no explicit control group or a biased one (not a random sample of the population) [6]. Given the natural difference between the treatment and control groups, and also the natural difference between consumption in 2012 and 2013, I use synthetic control to create a control group whose 2012 baseline is similar to that of the treatment group.

The treatment and control groups are said to both be representative samples of London as a whole as figures 3-3 and 3-2 show. It is, therefore, expected that the consumption for the treatment and control groups are similar in 2012. Having a similar baseline in 2012 is essential in being able to understand the treatment effect in 2013. However, given the double opt-in nature of this trial, the samples aren't random and there's selection bias in the treatment group. For this reason, finding the counterfactual and the treatment effect will require further analysis. Next, I'll be looking at just how different the behavior of the two groups were in the pre-intervention period.

3.3.1 Limitations and Normalizing the Data

If the treatment and control groups were random samples of households in London, given that they both went through the same static pricing scheme and general weather patterns during 2012, one would expect that once normalized for the differing size, that the two groups would have a similar consumption profile in 2012. Given that the trial is double opt-in, such expectation won't necessarily be true. Figures 3-4 and 3-5 show average consumption per day and average cost per day in 2012 and 2013. The difference between the average behavior of the two groups in the pre-intervention period demonstrates a natural difference between the populations of the dToU and nonToU groups. Therefore, further normalization and processing is needed, which will be described in chapter 4. The opt-in nature of the treatment group also limited the size of the treatment group to be smaller than that of the control group (roughly a ratio of 4:1) which can be resolved by looking at mean metrics when comparing across the two groups.

Though 3-4 and 3-5 demonstrate the necessity for normalizing to correct for the natural difference between the control and treatment groups, another way to estimate the treatment effect is to look at the percent difference between the two years within the same group. That analysis ignores the natural differences between the two years. In other words, even though we're comparing within the same group, the temperature and general circumstances change year-over-year. Figure 3-6 shows the mean consumption and mean cost over all houses present in the treatment and control groups over high, normal, and low hours, as well as all hours of the day. I should emphasize that this is without any normalization and to get a sense of how the treatment changed the behavior of the groups as compared to themselves the previous year, before the treatment.

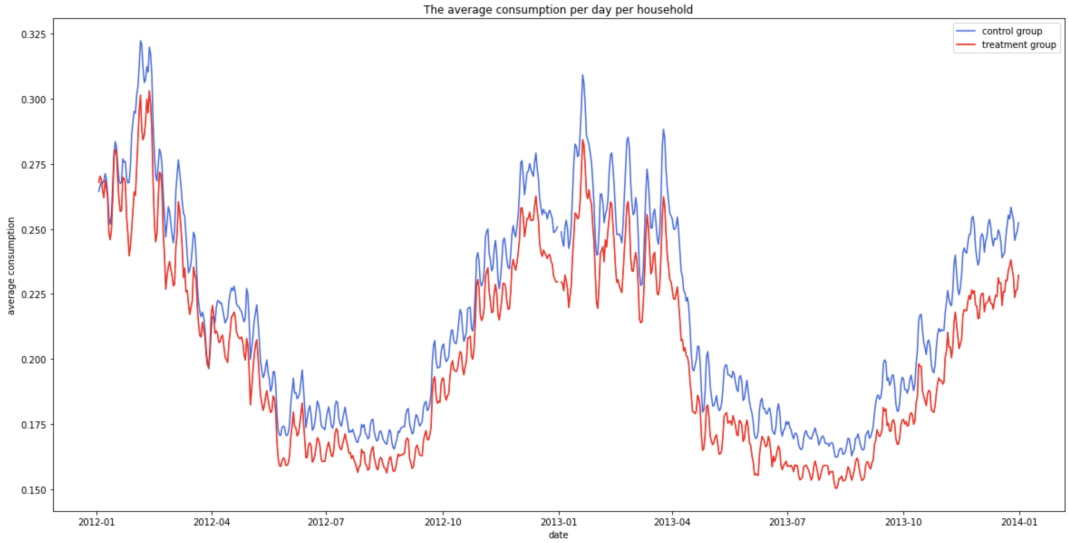

Figure 3-4: Average consumption per day per group. This shows that the houses have a fundamentally different consumption pattern even in 2012, the year the two groups had identical circumstances. This is a result of the fact that the treatment group opted-into being subject to dToU, which created a self-selecting treatment group, whose habits may have already differed (e.g. they were more energy conscious at the outset).

Comparing the percent difference in mean consumption, we can see that there's an increase in the consumption in the low hours for the treatment group (3.14%) and a decrease in the normal (-5.47%) and high hours (-9.05%). Similarly, comparing the percent difference in mean cost, we can see that there's a significant decrease in the cost for low hours for the treatment group (-71.03%) and a significant increase in the high hours (328.77%). This preliminary analysis is in line with the expected treatment effect; we expect consumption to be lower in high hours and higher in the low hours (shifted from the low hours) and for the cost to be higher during the high hours and lower during the low hours (even though consumption is expected to be high in the low hours) given the price bands for the treatment year.

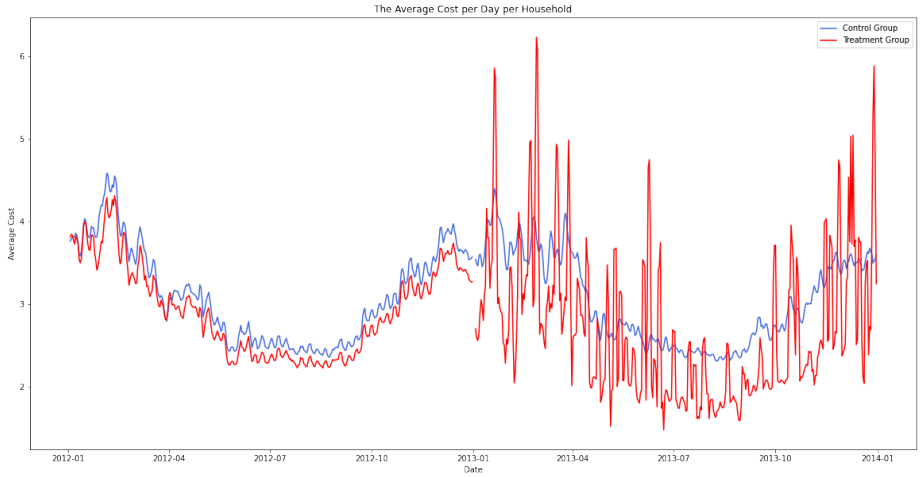

Figure 3-5: Average cost per day per group. Looking at 2012, as expected from 3-4, the treatment spent less on electricity (because they consumed less and 2012 had static electricity price). The fluctuations in 2013 are an indication that the treatment had a significantly higher cost during high price hours and lower in low price hours relative to the control group that was still experiencing static price.

3.4 Conclusion

In summary, equation 3.3 captures the problem at hand. Consumption in each year and for each group depends on temperature (T), time (t), and some group specific parameters (θ). There is a mapping between the two years (h) and another mapping between the two groups (g). The question is how best to use the control data from 2012/2013 and the treatment data from 2012 to find the counterfactual consumption for the treatment group in 2013.

$$\begin{array}{ccc} f_{2012}(\theta_{tr}, T, t) & \xrightarrow{h(t)} & f^*_{2013}(\theta_{tr}, T, t) \\ {\scriptstyle g(\theta_c)}\uparrow & & {\scriptstyle g(\theta_c)}\uparrow \\ f_{2012}(\theta_c, T, t) & \xrightarrow{h(t)} & f_{2013}(\theta_c, T, t) \end{array} \tag{3.3}$$

		Mean Consumption (kWh/hh)		Mean Cost (p/hh)	
		Treatment	Control	Treatment	Control
2012	Overall	0.207	0.22	2.94	3.12
	Low	0.191	0.212	2.72	3.02
	Normal	0.201	0.216	2.86	3.07
	High	0.243	0.253	3.46	3.60
2013	Overall	0.197	0.214	2.72	3.05
	Low	0.197	0.204	0.79	2.90
	Normal	0.19	0.209	2.24	2.98
	High	0.221	0.253	14.85	3.60
Delta	Overall	-4.83%	-2.73%	-7.51%	-2.46%
	Low	3.14%	-3.77%	-71.03%	-3.91%
	Normal	-5.47%	-3.24%	-21.66%	-3.09%
	High	-9.05%	0.00%	328.77%	0.06%

Figure 3-6: This table shows the mean consumption per group in (kWh/hh) and mean cost of electricity per group in (pence/hh) as well as the percentage difference between the two years within the same group. Without normalizing the control and treatment groups we can see that the treatment was effective.

Chapters 4 and 5 will outline in detail some models that find the counterfactual consumption using a mapping between the treatment and control groups in 2012. Chapters 4 will go through regression models that find a mapping between the 2012 treatment and control data. Chapter 5 includes a review of different time series models that can be used for a data set such as this and one implementation of a time series model with explicit dependence on time and temperature.

Chapter 4

Models & Results: Regression Models

This chapter includes different regression models for estimating the counterfactual and the resulting treatment effect, as well as the error on each model, the benefits, and shortfalls. I conclude the chapter with a comparison of the different models.

The chapter begins with a mathematical framing of the problem, it continues with an aggregated linear regression model and the resulting estimated treatment effect. The aggregated linear regression model aggregates values over all houses. Adding accuracy to that model, the multi linear regression model does the prediction without any aggregation on the household data and predicts on a per half-hour, per house level. Next is the aggregated multi linear regression model which falls in between the first two models in terms of aggregation — it aggregates over clusters of houses.

The chapter concludes with a future line of inquiry, a constrained optimization model which adds a constraint to the multi linear regression — that counterfactual consumption values should be positive, as well as a comparison of the different

models.

4.1 Research Question: Mathematical Framing of the Problem

Summarizing from 3.3, the question I will be answering here is whether the time-of-use pricing scheme carried out in the trial in 2013 was effective, on average, in lowering consumption during the high price hours. That will be the main question in focus. However, there will be other analysis around how the treatment affected consumption overall. All analyses is done in aggregate for a cluster of houses and not on a per house basis — more information on the reason behind this choice in section 4.3.2.

As mentioned in section 3.3, the goal is to find the treatment effect. The challenge with that goal is that I don't have a baseline consumption for the treatment group — I don't know what they would have consumed in 2013 in the absence of the treatment. Below I mathematically define some variables to quantify finding the counterfactual consumption.

Consider the following segmentation of the data. α and β are both matrices of

Year	Control Group	Treatment Group
2011	α_{2011}	β_{2011}
2012	α_{2012}	β_{2012}
2013	α_{2013}	$\beta_{2013}{}^{*}$
2014	α_{2014}	β_{2014}

dimension $t \times n$ where t is the number of time indices for that particular year ($= 365 \times 48 = 17520$ for year 2012 and 2013) and n is the number of houses present in each group during that year. The α and β matrices can be divided into different

socio-economic groups based on the `house_id` of the households. In that case, the consumption matrix only includes consumption values for households in a particular socio-economic group for a particular year. For example, the dimensionality gets reduced down to $t \times n_{\text{affluent}}$ as a result. This will become handy in the future sections as a mapping between specific socio-economic groups is explored.

The treatment group only went through the treatment in 2013. However, I have consumption data for this group when they were subject to static pricing in 2011, 2012, and 2014. Though I have consumption data for some houses starting in late 2011 and ending in February 2014, I have data for full calendar years of 2012 and 2013. As a result, the data from 2012 and 2013 is used in this analysis as it contains information regarding the effects of seasonal changes on energy consumption. 2012 is a leap year but Feb 29, 2012 was removed to allow for a 1-1 mapping between 2012 and 2013.

Additionally, note that these matrices will have missing values. In particular, α_{2012} and β_{2012} are going to be very sparse as a large number of the houses weren't recruited until later in the year as shown in figure 3-1. Consider $C_{m,h}$ the consumption of house h at time index m. Each element in the matrix is a half-hour consumption value for house h; i.e. $\alpha_{2012}^{m,h}$ occupies the m^{th} row and h^{th} column of the α_{2012} matrix and holds the consumption value of house h at time index m, $C_{m,h}$. Consider $\hat{\beta}_{2013}$ to be the counterfactual consumption for the treatment group in 2013 — what they would have consumed in the absence of the dynamic pricing scheme. To find whether the treatment was effective, I want to look at $\beta_{2013} - \hat{\beta}_{2013}$. Below are the different methods of finding $\hat{\beta}_{2013}$: aggregated linear regression model, multi linear regression, and aggregated multi linear regression.

4.2 Aggregate Linear Regression Model

As explained in more detail in section 3.1, due to the double opt-in nature of the trial, I need to find a relationship between the two groups using the data in 2012. I then apply that relationship to the data from 2013 to find the counterfactual consumption. In this section, I find a linear relationship between mean consumption values over all houses per time index. In this model, by taking the mean over all households, the α and β matrices go from consumption matrices with dimension $t \times n$ to vectors with dimension t, $\overline{\alpha^m}$ and $\overline{\beta^m}$. m here shows the time index and goes from 1 to t, the size of the vector.

I take the mean consumption value per time index for two reasons: 1. to normalize for the differing number of houses in each group[1] and 2. to normalize for missing data at the beginning of 2012: some houses were not present in the first months of 2012 as shown in figure 3-1.

As shown in equation 4.1 I'm regressing the treatment group's consumption on the control group's consumption, averaged over households to find the estimated counterfactual consumption, $\overline{\hat{\beta}^m_{2013}}$.

$$\begin{aligned} \overline{\beta^m_{2012}} &= a \times \overline{\alpha^m_{2012}} + b \\ \overline{\hat{\beta}^m_{2013}} &= a \times \overline{\alpha^m_{2013}} + b \\ \overline{\Delta\text{treatment}} &= \overline{\beta^m_{2013}} - \overline{\hat{\beta}^m_{2013}} \end{aligned} \tag{4.1}$$

$\overline{\alpha^m}$ and $\overline{\beta^m}$ are both vectors of size $t = 365 \times 48 = 17520$ for both 2012 and 2013. a and b are scalars that can be found from the regression on the aggregated 2012 data. I then apply the same regression to the 2013 control data to find the

[1] the treatment group is 1,100 households but the control group is of 4,467 households.

counterfactual consumption for the treatment group in 2013. The following figures show the 2012 control and treatment data, aggregated over all houses in figure 4-1 and aggregated per socio-economic group in figure 4-2.

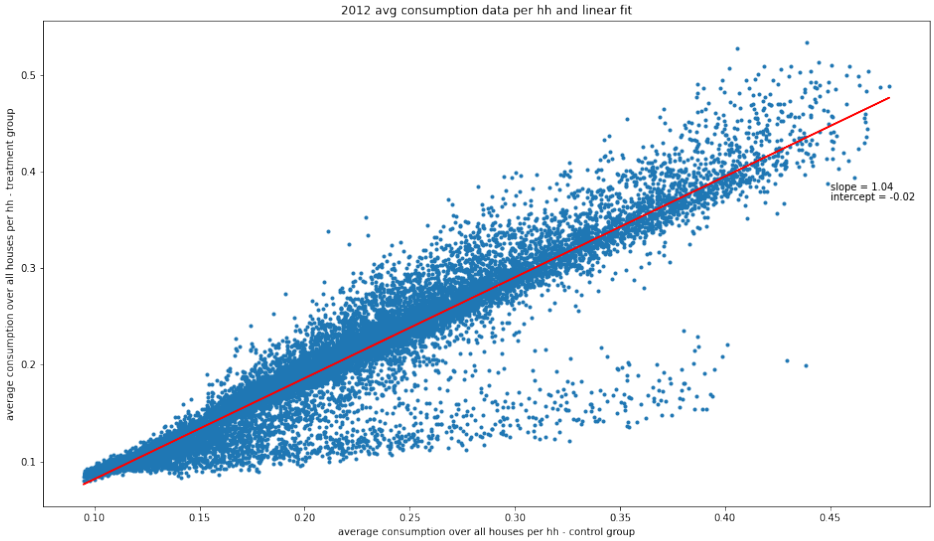

Figure 4-1: The linear mapping between the aggregate control group consumption per half-hour and the control group.

4.2.1 Error Analysis

The mean percent error on this method, as calculated by finding a linear mapping for 70% of all of the 2012 and 2013 control data and finding the error on the remainder is -2.58%. The mean error is -0.0001 and the standard deviation of error is 0.035. Table 4.1 shows the error on the mapping per socio-economic group.

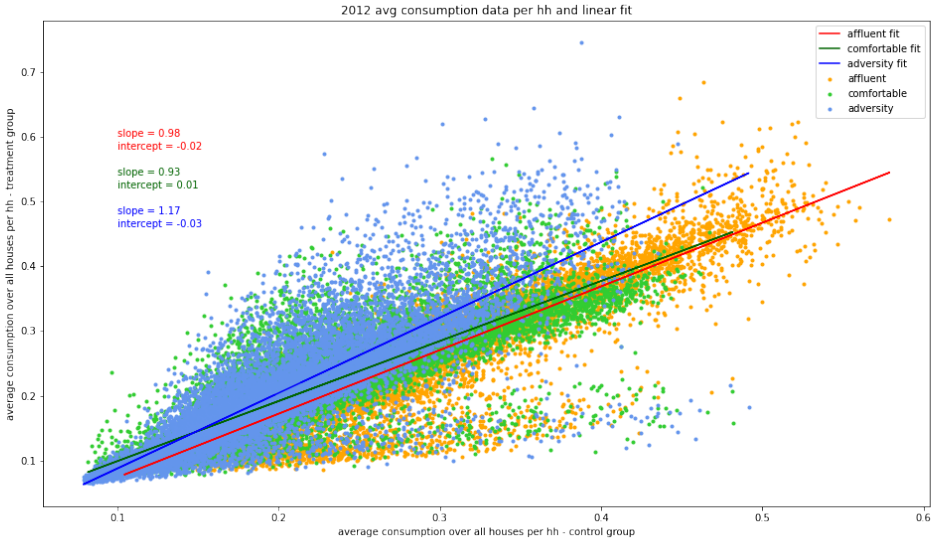

Figure 4-2: Aggregate linear regression between control group and treatment group in 2012, segmented by socio-economic status.

Aggregated Linear Regression Error Analysis			
Model	MPE	ME	SD
Aggregated linear regression (overall)	-2.58%	-0.0001	0.035
Aggregated linear (affluent)	-2.68%	00.0002	0.0421
Aggregated linear (comfortable)	-3.92%	-0.0003	0.0422
Aggregated linear (adversity)	-3.38%	-0.0002	0.0323

Table 4.1: Error analysis for the aggregated linear regression model used on the control group reported in mean percent error (MPE), mean error (ME), and standard deviation of error (SD). The training size is 0.7 in all cases.

4.2.2 Counterfactual Analysis

Referring back to 4.1, I need to find a and b to be able to find the mean counterfactual consumption $\overline{\hat{\beta}^m_{2013}}$. a and b are scalars whose values can be found in figures 4-1 and 4-2. It's important to note that the specific values are not important. What matters

most the estimated counterfactual consumption, $\overline{\hat{\beta}^m_{2013}}$, and more importantly, the resulting treatment effect, $\overline{\beta^m_{2013}} - \overline{\hat{\beta}^m_{2013}}$ which can be found by applying those values to the 2013 control data as shown in equation 4.1.

Table 4-3 shows the mean and mean percentage difference between the actual data and counterfactual consumption as found through the linear mappings shown above. All the values on right hand side are found through mapping per socio-economic group. The values on the left hand side are found through a single mapping over all houses. The first and second rows compare the effect of the treatment on consumption. The first row of the table shows the mean treatment effect in kWh/hh. The second row of the table shows the percent treatment effect, so the change relative to what the actual consumption was. The third and fourth rows compare the effect of the treatment on electricity cost. The third row shows the effect on the treatment in pence/hh. The fourth row shows the percent treatment effect i.e. the percent change in cost compared to what the actual cost was in 2013.

Table 4-3 shows that all socio-economic groups shed demand during the high hours. The mapping that averages over all the houses in 2012 (the tables on the left hand side) shows that the affluent group shed demand during all hours of the day whereas the comfortable and adversity groups consumed more during the low hours. The mapping that averages over houses within a single socio-economic group (the tables on the right hand side) shows the opposite, that the adversity group shed consumption during all hours, most in high hours, next in normal hours, least in low hours. The treatment affected the comfortable group similarly per this linear mapping. The affluent group shed consumption during the high hours (less than other socio-economic groups) and increased demand in low and normal hours. Even though there is a discrepancy here on which socio-economic group shed most, both

Mean Error Between Counterfactual and Real Consumption for 2013 (regression on all of 2012)					Mean Error Between Counterfactual and Real Consumption for 2013 (regression on socio-economic groups)				
	Low	Normal	High	Overall		Low	Normal	High	Overall
Affluent	-0.0084	-0.0160	-0.0345	-0.0161	Affluent	0.0085	0.0005	-0.0149	0.0005
Comfortable	0.0044	-0.0097	-0.0326	-0.0094	Comfortable	0.0004	-0.0141	-0.0314	-0.0135
Adversity	0.0134	0.0024	-0.0085	0.0029	Adversity	-0.0022	-0.0129	-0.0281	-0.0126
Overall	0.0059	-0.0046	-0.0209	-0.0043	Overall	0.0059	-0.0046	-0.0209	-0.0043
Mean Percent Error Between Counterfactual and Real Consumption for 2013 (regression on all of 2012)					**Mean Percent Error Between Counterfactual and Real Consumption for 2013 (regression on socio-economic groups)**				
	Low	Normal	High	Overall		Low	Normal	High	Overall
Affluent	-5.24%	-7.48%	-13.44%	-7.54%	Affluent	2.47%	0.49%	-5.40%	0.41%
Comfortable	2.21%	-3.80%	-13.24%	-3.66%	Comfortable	-1.84%	-8.39%	-14.76%	-8.06%
Adversity	5.96%	1.16%	-4.69%	1.35%	Adversity	-2.63%	-7.83%	-14.71%	-7.65%
Overall	1.87%	-1.97%	-8.60%	-1.90%	Overall	1.87%	-1.97%	-8.60%	-1.90%
Mean Error Between Counterfactual and Real Cost for 2013 (regression on all of 2012)					**Mean Error Between Counterfactual and Real Cost for 2013 (regression on socio-economic groups)**				
	Low	Normal	High	Overall		Low	Normal	High	Overall
Affluent	-2.4650	-0.7570	12.8927	-0.3049	Affluent	-2.2247	-0.5224	13.1726	-0.0678
Comfortable	-2.0450	-0.6031	10.9730	-0.2191	Comfortable	-2.1019	-0.6652	10.9901	-0.2771
Adversity	-1.6267	-0.3721	9.9642	-0.0260	Adversity	-1.8479	-0.5899	9.6850	-0.2470
Overall	-2.0525	-0.5433	11.6571	-0.1376	Overall	-2.0525	-0.5433	11.6571	-0.1376
Mean Percent Error Between Cost and Real Consumption for 2013 (regression on all of 2012)					**Mean Percent Error Between Counterfactual and Real Cost for 2013 (regression on socio-economic group)**				
	Low	Normal	High	Overall		Low	Normal	High	Overall
Affluent	-275.28%	-30.04%	75.98%	-48.51%	Affluent	-247.78%	-20.39%	77.68%	-37.53%
Comfortable	-248.72%	-25.59%	76.02%	-42.16%	Comfortable	-263.16%	-31.14%	75.70%	-48.32%
Adversity	-235.34%	-19.59%	77.83%	-35.65%	Adversity	-265.96%	-30.46%	75.71%	-48.00%
Overall	-249.93%	-23.37%	77.01%	-40.32%	Overall	-249.93%	-23.37%	77.01%	-40.32%

Figure 4-3: This table shows the mean and mean percentage difference between the counterfactual and real consumption of the treatment group in 2013. The counterfactual is calculated both using the linear mapping from regression on all of 2012 and regression per socio-economic status. In this heat map, the smaller values are red and the larger values are green.

estimates show that all socio-economic groups responded to the treatment effect. The results from the linear mapping per socio-economic (the table on the right) are more accurate as this regression is tighter on the consumption of per socio-economic group. For this reason, the results from the right hand side are intuitively more trust-worthy.

Both regressions show similar results in terms of the effect of the treatment on average cost of electricity during low, normal, and high hours. The treatment shed ∼250% from low hours expenditure, ∼23% from normal hours expenditure, and increased expenditure in high hours by ∼77%. This is a percentage difference compared to the expenditure in those hours in the absence of the treatment. Even though consumption was higher in the low hours, the -10.238 p/kWh difference in tariff in low hours makes the ∼250% decrease possible. Similarly, though consumption was lower in the high hours, the 52.972 p/kWh difference in cost during those hours is reason for the ∼77% increase on expenditure during these hours.

Table 4-4 shows the standard deviation of the distribution of treatment effects.

STD of Error Between Counterfactual and Real Consumption for 2013 (regression on all of 2012)					**STD of Error Between Counterfactual and Real Consumption for 2013 (regression on socio-economic groups)**				
	Low	**Normal**	**High**	**Overall**		**Low**	**Normal**	**High**	**Overall**
Affluent	0.0237	0.0182	0.0241	0.0196	**Affluent**	0.0240	0.0163	0.0203	0.0178
Comfortable	0.0236	0.0189	0.0273	0.0209	**Comfortable**	0.0231	0.0142	0.0191	0.0165
Adversity	0.0220	0.0139	0.0170	0.0155	**Adversity**	0.0213	0.0160	0.0210	0.0174
Overall	0.0195	0.0119	0.0176	0.0139	**Overall**	0.0195	0.0119	0.0176	0.0139
STD of Error Between Counterfactual and Real Cost for 2013 (regression on all of 2012)					**STD of Error Between Counterfactual and Real Cost for 2013 (regression on socio-economic groups)**				
	Low	**Normal**	**High**	**Overall**		**Low**	**Normal**	**High**	**Overall**
Affluent	0.8778	0.3831	4.9640	3.1240	**Affluent**	0.8103	0.3202	5.0514	3.1308
Comfortable	0.8403	0.3909	3.7469	2.6280	**Comfortable**	0.7171	0.2792	3.8947	2.6368
Adversity	0.5901	0.2273	3.5881	2.3433	**Adversity**	0.6890	0.3073	3.4750	2.3344
Overall	0.7516	0.2837	4.2264	2.7699	**Overall**	0.7516	0.2837	4.2264	2.7699

Figure 4-4: This table shows the standard deviation of the difference between the real consumption and cost of the treatment group and the estimated counterfactual. The table on the left shows that value as estimated by the linear mapping between an aggregation of all the houses and the table on the right.

4.2.3 Hypothesis Testing

Going back to the hypotheses in section 3.2, we want to test two things.

First, we want to run a one-sided t-test to see if the mean treatment effect in the high price hours is less than zero. If the p-values are statistically significant, it would mean that the treatment successfully lowered consumption during the high price hours, which was the goal of the dToU mechanism. Table 4.2 shows p-values resulting from this t-test. In order to find the p-values shown, I run a `ttest_rel` using the `scipy.stats` package on the vectors of actual consumption and the estimated counterfactual consumption. The t-test is run to detect whether the mean of the distribution underlying the actual consumption sample is less than the mean of the distribution underlying the counterfactual consumption sample.

Socio-economic Group	Affluent	Comfortable	Adversity	Overall
p-value (overall mapping)	7.3e-193	5.9e-154	1.8e-40	1.9e-152
p-value (per socio-economic mapping)	8.3e-76	2.2e-226	4.0e-178	`N/A`

Table 4.2: The p-values to test the null hypothesis that the ATE in the high hours is less than zero, for both the overall and per socio-economic mapping.

These p-values show strong evidence against the null hypothesis and prove that the treatment was effective in lowering consumption during the high price hours in every socio-economic group.

Second, we want to show that the adversity socio-economic group is more price sensitive than the comfortable socio-economic group and the comfortable socio-economic group is more price sensitive than the affluent socio-economic group. For this purpose, I will run two one-sided paired t-tests, one to show that ATE(affluent) $-$ ATE(comfortable) > 0 and ATE(comfortable) $-$ ATE(adversity) > 0. It is important to note that the treatment effect here is the effect of the treatment on the **cost of electricity**. We are no longer conditioning on any price hour as we want to see overall, how the different socio-economic groups responded to the treatment.

I similarly run a `ttest_rel` using the `scipy.stats` package on the vectors of counterfactual cost per socio-economic group. The one-sided t-test is such that the mean of the counterfactual cost distribution in the affluent group is more than the mean of the counterfactual cost distribution in the comfortable group and the mean of the counterfactual cost distribution in the comfortable group is more than the mean of the counterfactual cost distribution in the adversity group groups. The t-test was run on on the counterfactual cost found by both the overall mapping and the per socio-economic mapping. All four p-values are zero and show that we can reject the null. Socio-economic status affects the response to the treatment. Additionally, the one-sided t-test proves that the the mean of the counterfactual cost distribution in the affluent group is more than the mean of the counterfactual cost distribution in the comfortable group and the mean of the counterfactual cost distribution in the comfortable group is more than the mean of the counterfactual cost distribution in the adversity group groups.

4.2.4 Limitations and Conclusion

This model aggregates over all data points and finds a linear mapping between the control group and the treatment group which arguably loses a large amount of information present in the data set. Additionally, given that the model involves both the control and treatment groups, the method with which I found the error is not fundamentally what the model is built to capture. In other words, by finding a linear mapping between 70% of the 2012/2013 control data and finding the error on the remaining data, we're implicitly assuming a linear mapping between the same group over the two years, which is not what the model assumes. The model assumes a linear relationship between the consumption behavior of the two groups.

4.3 Multiple Linear Regression Model

In the aggregate linear regression model outlined in section 4.2, I took the mean over all household consumption values, effectively going from consumption matrices with dimension $t \times n$ to vectors with dimension t. Arguably, there is important information within those data points that is lost through applying an aggregate function. Another approach is to commit to using every data point present in the data set to find the counterfactual. It might be the case that there are patterns in household behaviors outside of the pre-defined clusters (socio-economic status). This multi linear regression model finds a mapping per house per half-hour using all data points.

In order to find $\hat{\beta}_{2013}$, let us assume a relationship between the two groups that can be captured by $\alpha_{2012}X = \beta_{2012}$. Since α_{2012} is not an invertible matrix, I find its inverse using the Moore-Penrose inverse. $X = \alpha_{2012}^{-1}\beta_{2012}$ captures the way the two groups map to one another. Assuming that the nature of the control and treatment groups stay the same throughout the trial, I can find the counterfactual consumption for the treatment group with the same mapping, i.e. $\hat{\beta}_{2013} = \alpha_{2013}X$. To summarize, the model is as follows:

$$\begin{aligned} \alpha_{2012}X &= \beta_{2012} \\ \alpha_{2013}X &= \hat{\beta}_{2013} \\ \Delta\text{treatment} &= \beta_{2013} - \hat{\beta}_{2013} \end{aligned} \tag{4.2}$$

4.3.1 Matrix Imputation

Given equation 4.2, the dimensionality of the consumption matrices must match across 2012 and 2013 i.e. the α_{2012} and α_{2013} are of the same dimension and β_{2012} and β_{2013} are of the same dimension. This, however, poses a constraint as much of the 2012 matrices will be empty (recall that the recruitment was ongoing so much of the 2012 data is missing for the first half and for houses that were recruited later in 2012). Additionally, given that my goal for this model was to keep all and use all the available data, I decided to fill out the data points that don't exist in the data set.

The imputation function takes the median consumption value for a particular time index across all houses present in the α or β matrix and fills in the missing consumption values for that time index with that value. This means that if the analysis is being done on a subset of the data — for example, on a particular socio-economic status — the median is taken across all houses in that subset.

The imputed values are simply an approximation of how much households with missing data might have consumed in 2012. For this reason, I decided to find how much error imputing the same median value for all houses would introduce. In order to quantify the error introduced, I used a subset of the 2013 control data (that had data present for the entire year), masked a fraction of the data, and found the error on the imputed values. Figure 4-5 shows the amount of error introduced vs percentage of data imputed.

Figure 4-6 shows the percentage of the data missing across the time index axis in 2012 and 2013. Since α_{2012} and β_{2012} have a large fraction of the time index missing for the first six months, I limited my analysis to the latter six months. Recall that the trial runs until February 2014 which means that there exists two months of out

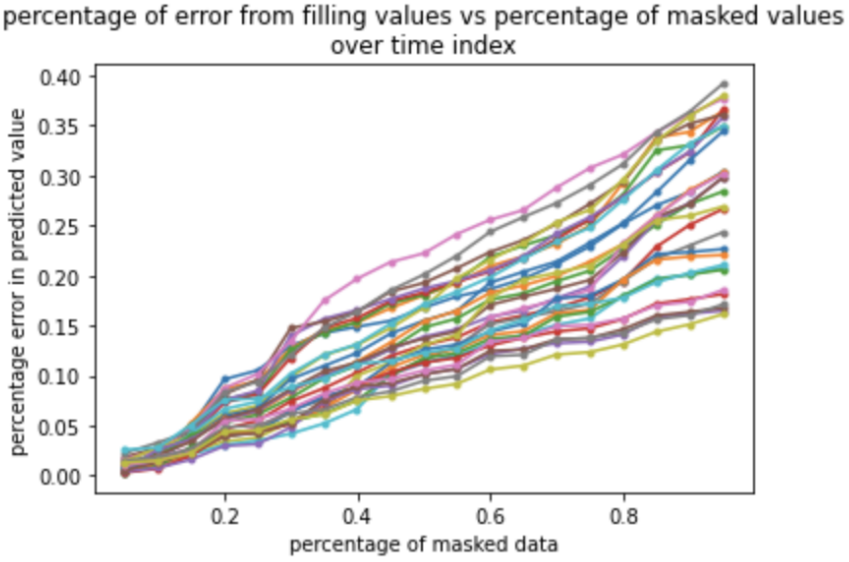

Figure 4-5: To understand the amount of introduced error by imputing consumption values, I masked some fraction of the data and then found the error introduced. This plot shows the error vs the percentage of the masked data.

of sample data for the control and treatment groups. 2014 has far fewer missing data — figure 4-7 shows the percentage of time index data missing in the first two months of 2013 and 2014. This data can be used to get another estimate for the treatment effect for the January and February of 2013 as shown in equation 4.3 where the α and β matrices are sliced to only include data from January and February.

$$\begin{aligned} \alpha_{2014} X &= \beta_{2014} \\ \alpha_{2013} X &= \hat{\beta}_{2013} \\ \Delta\text{treatment} &= \beta_{2013} - \hat{\beta}_{2013} \end{aligned} \tag{4.3}$$

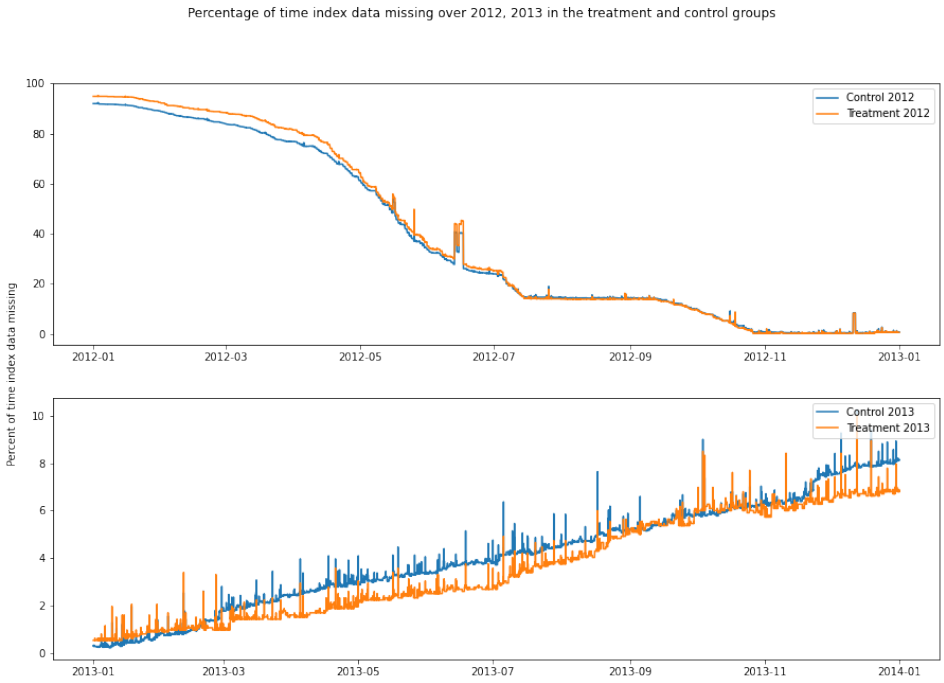

Figure 4-6: The percentage of time index data missing in α_{2012}, β_{2012}, α_{2013}, and β_{2013}.

4.3.2 Error Analysis

This multi linear regression model finds the counterfactual consumption per house per half-hour. In order to quantify the error on this method, I trained the model on a percentage of the control group's data and calculated the error on the remaining data for the control group. Given that the control group has never gone through any treatment, all the data is real and hence, the error can accurately be calculated.

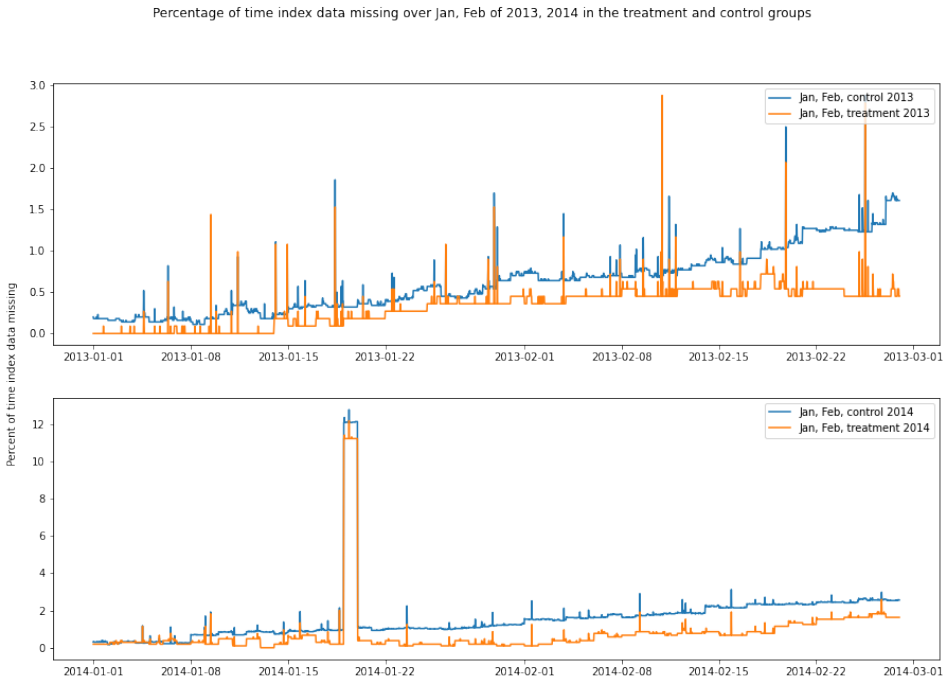

Figure 4-7: The percentage of time index data missing in January and February of 2013 and 2014.

$$\begin{aligned} \text{train}_{2013}X &= \text{train}_{2014} \\ \text{test}_{2013}X &= \hat{\text{test}}_{2014} \end{aligned} \tag{4.4}$$

The train matrices, train_{2013} and train_{2014} are both of dimension $N_{train} \times t$ where N_{train} are the houses in the the training set. They number of houses in both the 2013 and 2014 training set must be the same to make the dimensionality of equation 4.4 work. To make sure I have a mapping for the same subset of houses, I find the intersection of the houses in both the 2013 and 2014 control group to make the

number of houses in the training groups the same. For reasons I will expand on below, it's not necessary to find the intersection: simply making sure the number of houses is the same is sufficient. The data has been cleaned such that there is no missing data whatsoever — all houses with missing time index values were removed. This is due to the fact that I wanted to remove any change to the data that might introduce an error to isolate the error present in the model itself as much as possible. A similar equation as equation 4.4 can be written for the 2012/2013 control data.

Attempting to quantify the error here brings up a question: do I trust the values in the outcome matrix $\hat{\text{train}}_{2014}$ such that I want to calculate the deviations of every predicted value per half-hour per house? Or, does it make more sense to look at the predictions in some aggregate format?

Figures 4-8 and 4-9 show $\hat{\text{train}}_{2014}$ and train_{2014}, the predicted and real consumption values respectively, per house (4-8) and aggregated over a number of houses (4-9). As can be seen, the prediction is far from accurate when looked at on per house granularity. The prediction, and hence the error analysis, start to make sense when looked at in aggregate.

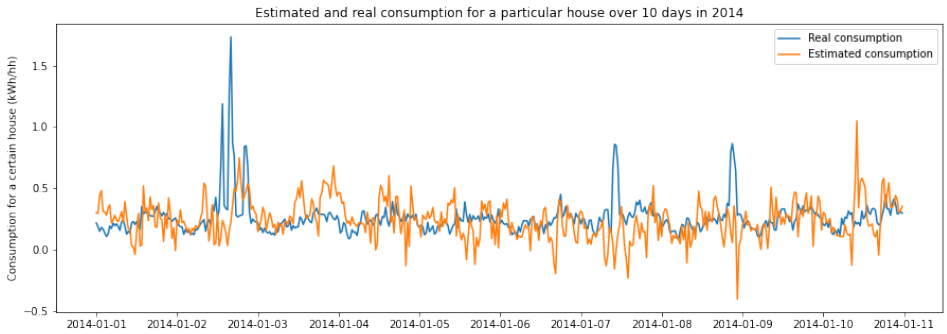

Figure 4-8: This shows the estimated and real consumption for a single house over 10 days. The prediction is far from the real consumption.

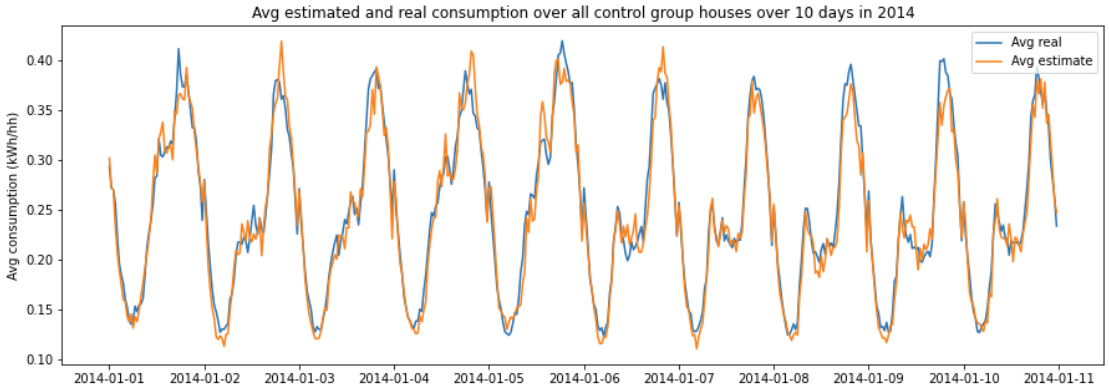

Figure 4-9: This shows the estimated and real consumption aggregated over all houses over 10 days. The prediction follows the real value closely when looked at in aggregation.

Even when comparing the real consumption of a single house to the mean consumption averaged over all houses, as figure 4-10 shows, a single household's consumption doesn't have a well-defined wave form whereas the aggregate has explicit trends.

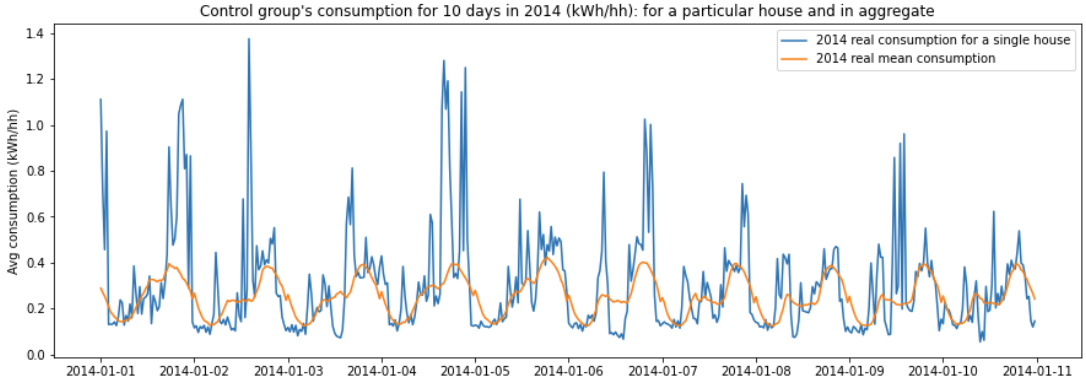

Figure 4-10: A single house's consumption over 10 days in 2014 vs. the aggregated consumption over all houses. The aggregated consumption demonstrates well-defined temporal trends.

In conclusion, when looking at the estimated counterfactual consumption and the

treatment effect, it is more meaningful to either predict in aggregate format (similar to the aggregate linear regression method) or to predict per half-hour per household, as this method does, and analyze values in aggregate. The per household prediction effectively has no meaning. This matches expectation as this method is not explicitly made to pull out patterns of consumption per household. It is for this reason that it's not necessary to take the intersection of the 2013 and 2014 train matrices, making the matrices have the same number of houses is sufficient. One could either take the intersection or drop some houses randomly to prevent any biases from being introduced.

One point to note here is that this method results in the counterfactual matrix $\hat{\text{test}}_{2014}$ whereas the aggregated linear regression method outlined in section 4.2 found the counterfactual consumption aggregated over houses and as a time series vector. One way to find the mean treatment effect or mean percent treatment effect in this method is to take the difference between the $\hat{\text{test}}_{2014}$ matrix and the test_{2014} matrix and then take the mean over households and then over the time index. Another way is to take the mean over households of both matrices first and find the error as I did in 4.2, by taking the difference of the two vectors and taking the mean of the resulting time series.

As figure 4-8 shows, the real consumption values are often very small compared to the estimated value for that household and time index. For this reason a mean percent error before taking the mean over all the households leads to incredibly large values. The two methods leads to the same mean treatment effect. Figure 4-11 shows how the mean percent error changes with training size. As expected, the error gets smaller as the size of the training sample gets larger. Given the conclusion drawn from figure 4-10, if the training size is too large, then the test sample will be too small and will not have enough data such that the average can smooth out the wobbles.

There is no imputed data in this entire analysis.

One question that may arise here is why are errors larger when the data is segmented by socio-economic group? The expectation would be that if the group is segmented with houses that are alike, the mapping would be tighter and more accurate. One possible explanation is there might be useful information across socio-economic groups that the model isn't able to learn or draw from when segmented. Along the same lines, the model benefits from a larger number of houses and when segmented per socio-economic group, there are fewer houses present. A less likely explanation is that the 2013/2014 data is too short (Jan 1 to Feb 27) for the model to capture the trends. It's possible to eliminate the latter explanation by rerunning the same analysis on the 2012/2013 control data which span over a longer period of time. Figure 4-12 shows the error in the same analysis run on the 2012/2013 data with time values ranging from July 1 to Dec 31. Of note is that given the missing values in the 2012 data, there are fewer houses that could be included in the 2012/2013 analysis and that was the reason behind picking 2013/2014 in the first place.

4.3.3 Counterfactual Analysis

Going back to figure 4-1 I found a mapping from mean consumption over all the houses in the control group to the mean consumption over all the houses in the treatment group. In figure 4-2 I found this mapping with aggregations done over all houses within a socio-economic status. I will do the same here i.e. the α and β matrices in equation 4.2 will include all houses when finding the treatment effect over all the houses and it will include the subset of houses in a particular socio-economic status when finding the mapping per socio-economic group. In addition, in this section, I will be running each of the above analyses both with imputed data

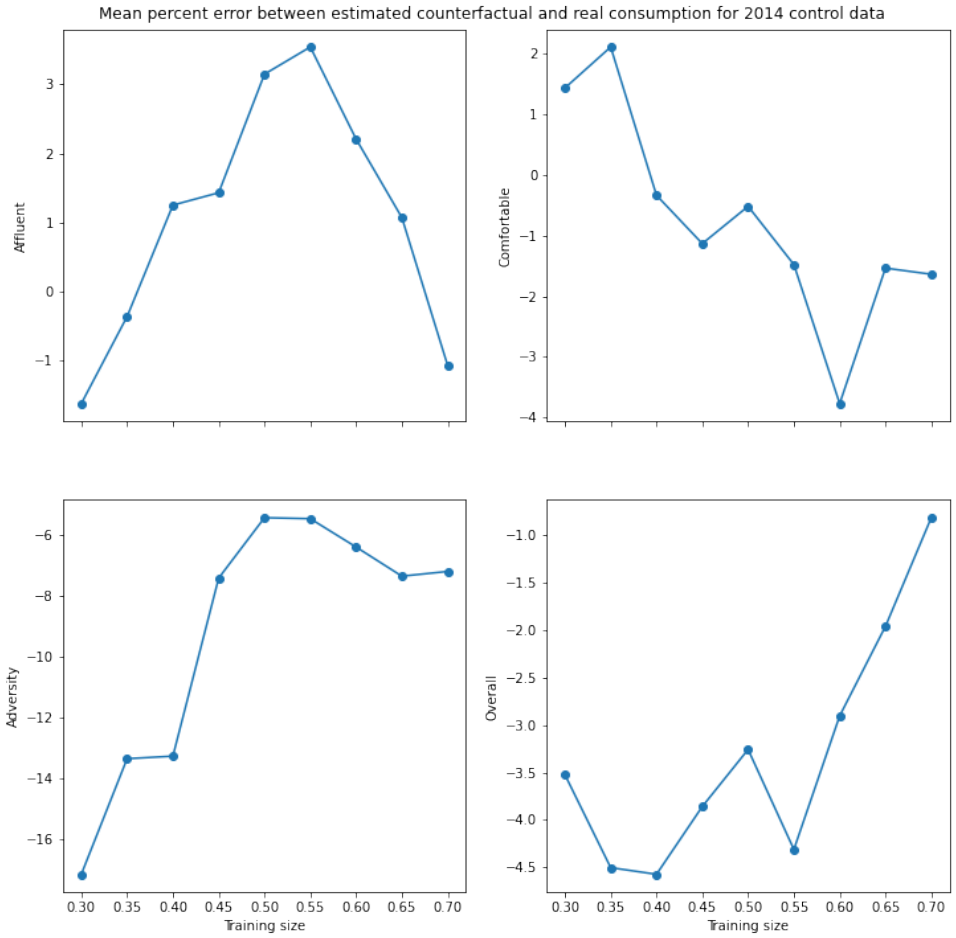

Figure 4-11: The mean percent error between the real consumption values and estimated counterfactual consumption on the 2014 test set. The mapping is found per socio-economic group and for the entirety of the data.

and without. That creates four different counterfactual analyses: mapping over all the houses (with and without imputation), mapping per socio-economic group (with

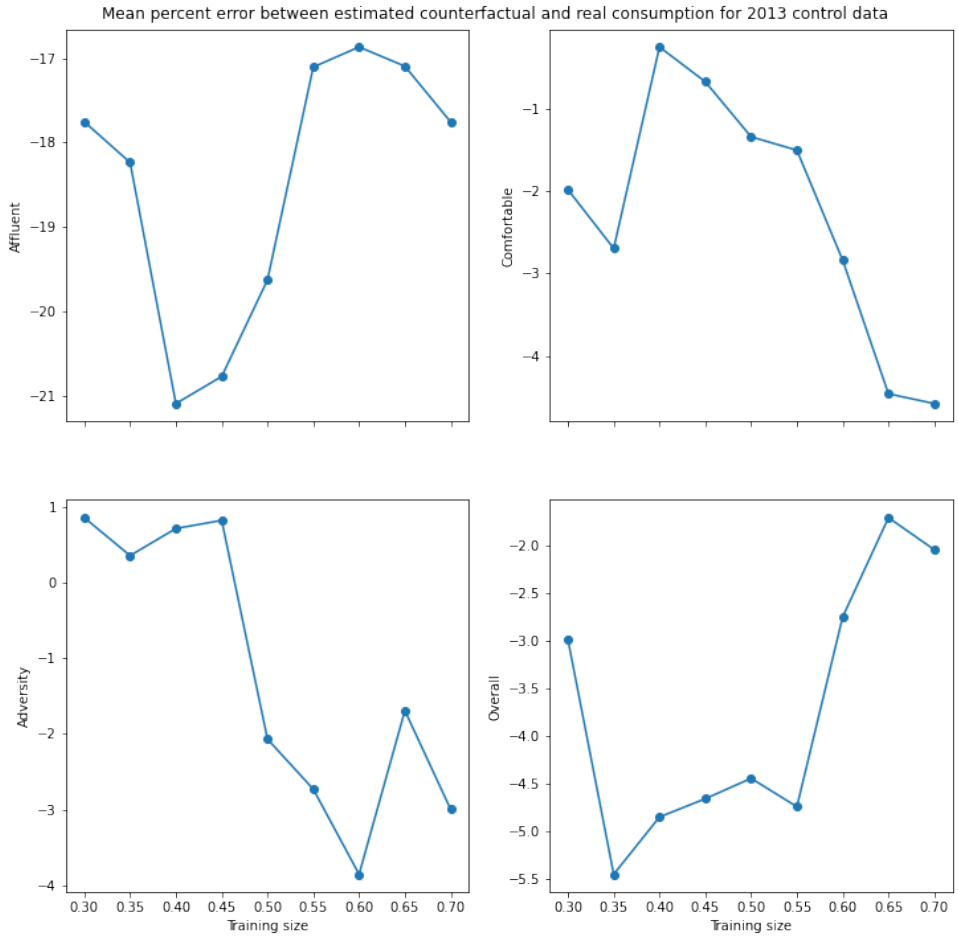

Figure 4-12: The mean percent error between the real consumption values and estimated counterfactual consumption on the 2013 test set. The mapping is found per socio-economic group and for the entirety of the data.

and without imputation).

First, I will be looking at treatment effect resulting from mapping all control

houses to treatment houses in 2012. In tables 4-13, 4-14, 4-15, and 4-16, the treatment effects are outlined both as mean error values (kWh/hh) and also in percent error (on average, how much of their consumption did households shed). Data from the second half of 2012/2013 is used for this analysis. This is due to the fact that there are fewer data points missing in the second half of 2012 as some houses were not recruited until later in 2012. This results in being forced to use data from the second half of 2013 as well, to make the dimensionality of equation 4.2 work.

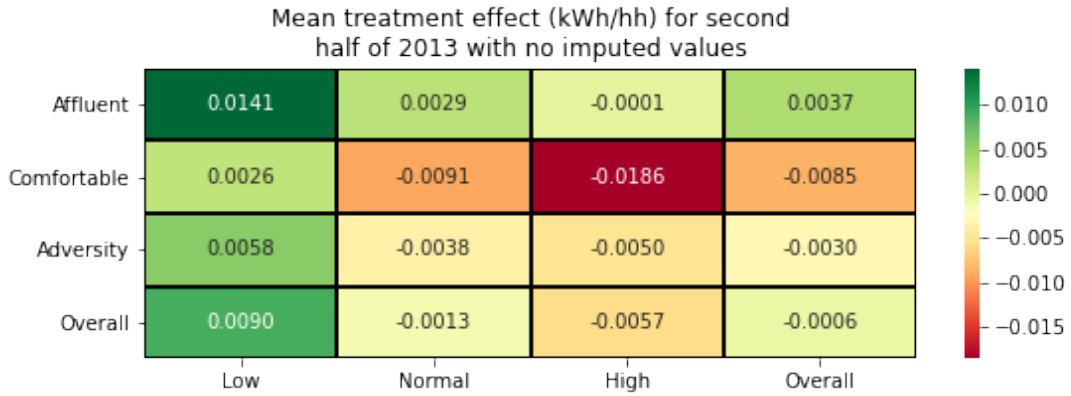

Figure 4-13: The mean treatment effect broken down per socio-economic status and hour of the day. The mapping is from all houses in the control group to all houses in the treatment group in the second half of 2012 and 2013. All houses with missing values have been dropped.

Table 4-14 shows the percent treatment effect with no imputed values; all houses with missing data were removed. It shows all socio-economic groups responded to the treatment and shed consumption during the high hours; the comfortable group most with 10.25% shed and the affluent group least with 3.44% shed. All groups similarly shed during the normal hours; the comfortable group most with 6.63% shed and the affluent group least with 0.09% shed. During the low hours, the comfortable group is found to have shed 1.6% on average whereas the other two socio-economic groups

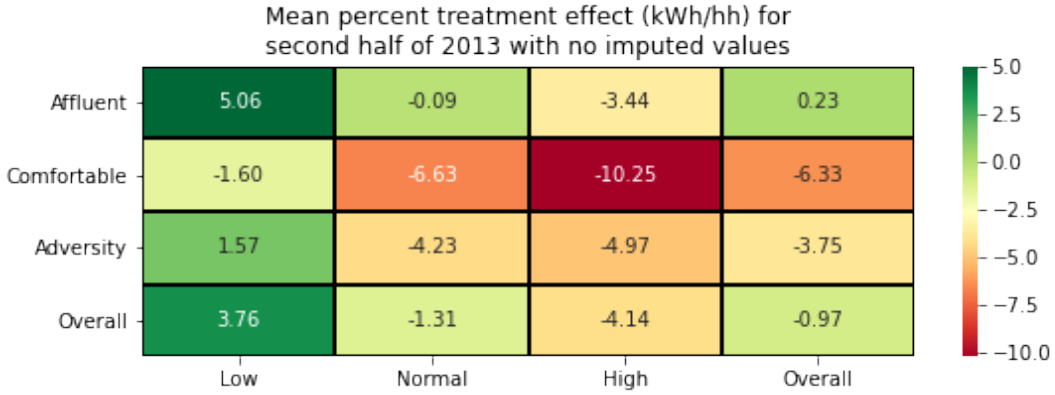

Figure 4-14: The mean percent treatment effect broken down per socio-economic status and hour of the day. The mapping is from all houses in the control group to all houses in the treatment group in the second half of 2012 and 2013. All houses with missing values have been dropped.

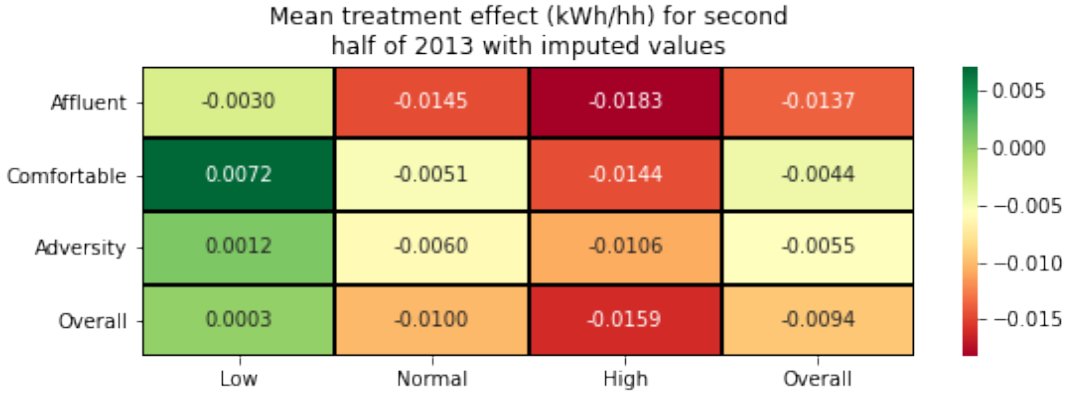

Figure 4-15: The mean treatment effect broken down per socio-economic status and hour of the day. The mapping is from all houses in the control group to all houses in the treatment group in the second half of 2012 and 2013 and for missing values imputed.

have increased consumption.

Table 4-16 shows the percent treatment effect with imputed values. It shows all socio-economic groups responded to the treatment and shed consumption during the

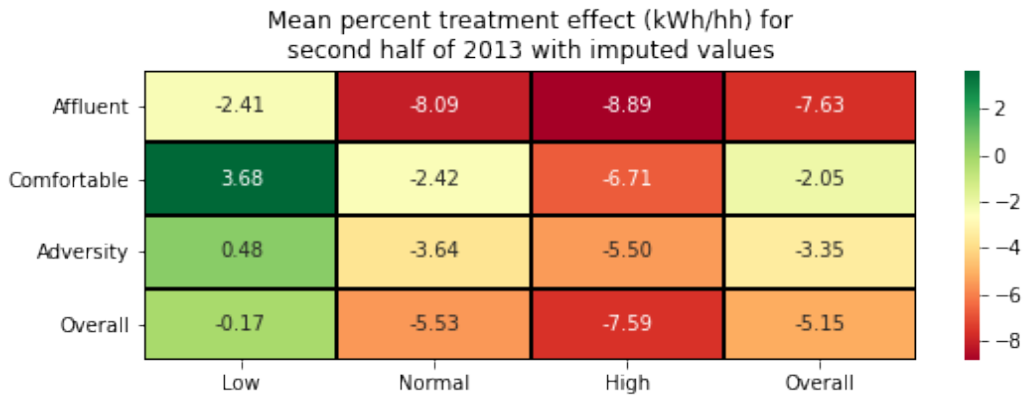

Figure 4-16: The mean percent treatment effect broken down per socio-economic status and hour of the day. The mapping is from all houses in the control group to all houses in the treatment group in the second half of 2012 and 2013 and for missing values imputed.

high hours; the affluent group most with 8.89% shed and the adversity group least with 5.5% shed. All groups similarly shed during the normal hours; the affluent group most with 8.09% shed and the comfortable group least with 2.42% shed. During the low hours, the affluent group is found to have shed 2.41% on average whereas the other two socio-economic groups have increased consumption.

The percent treatment effect found with missing values imputed or dropped lead to different results; they both show highest percent shed during the high hours overall, next during the normal hours, and least during the normal hours. This is in line with expectations and the goal for the time-of-day pricing. However, the results show different level of responsiveness to the treatment among the socio-economic groups. Table 4-14 shows the comfortable group shed most whereas table 4-16 shows the affluent group shed most. This will lead to different conclusions in terms of which group is most price sensitive as shown by their behavior in response to the time-of-day pricing. Before I make any conclusions on price sensitivity, I will be doing the

same analysis as above but with the mapping per socio-economic group.

The following tables includes the same analysis but with the mapping found on per socio-economic group.

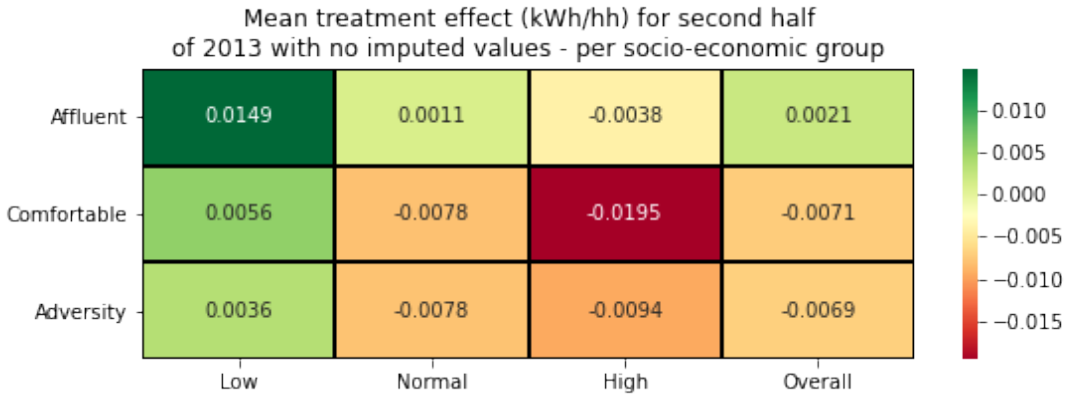

Figure 4-17: The mean treatment effect broken down per socio-economic status and hour of the day. The mapping is per socio-economic group in the second half of 2012 and 2013. All houses with missing values have been dropped.

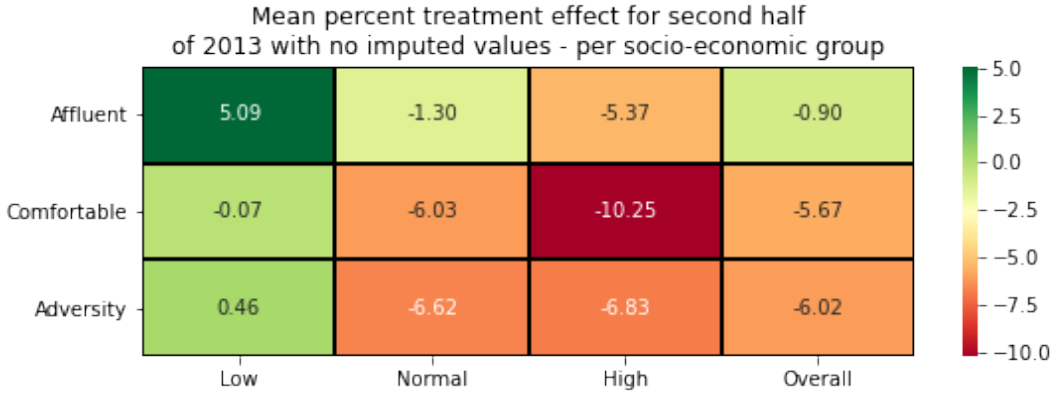

Figure 4-18: The mean percent treatment effect broken down per socio-economic status and hour of the day. The mapping is per socio-economic group in the second half of 2012 and 2013. All houses with missing values have been dropped.

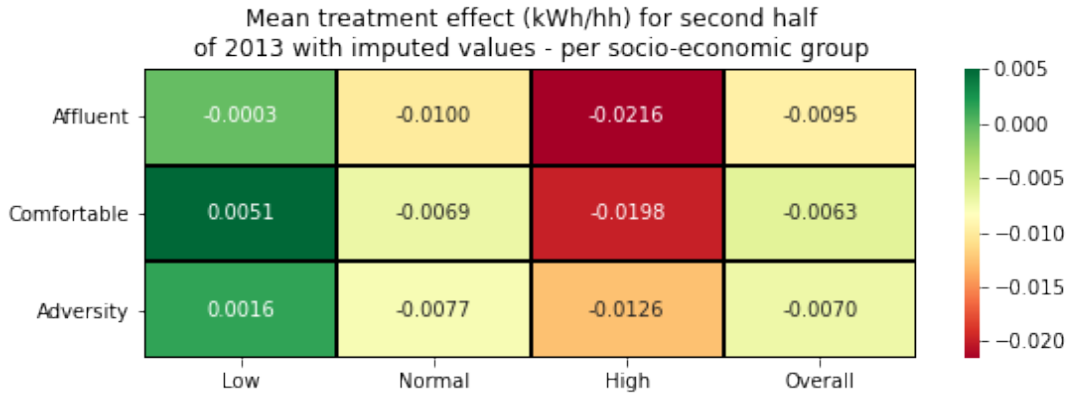

Figure 4-19: The mean treatment effect broken down per socio-economic status and hour of the day. The mapping is per socio-economic group in the second half of 2012 and 2013 and for missing values imputed.

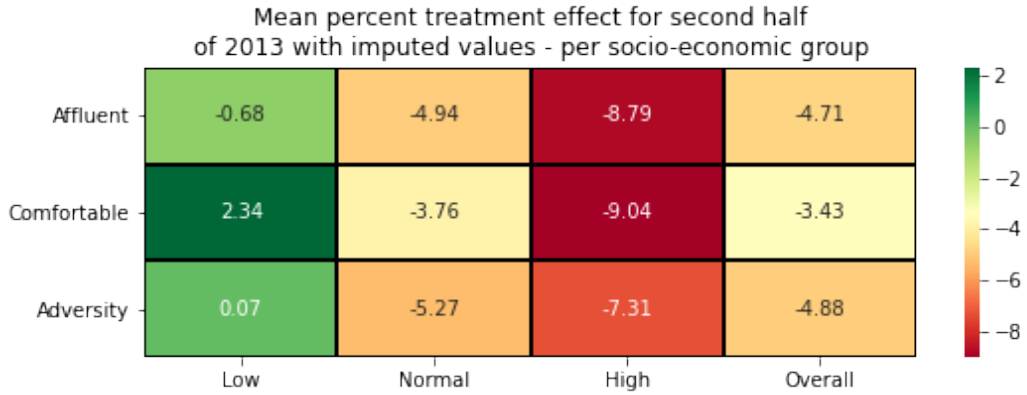

Figure 4-20: The mean percent treatment effect broken down per socio-economic status and hour of the day. The mapping is per socio-economic group in the second half of 2012 and 2013 and for missing values imputed.

Table 4-18 shows the percent treatment effect with no imputed values; all houses with missing data were removed. It shows all socio-economic groups responded to the treatment and shed consumption during the high hours; the comfortable group most with 10.25% shed and the affluent group least with 5.37% shed. All groups

similarly shed during the normal hours; the adversity group most with 6.62% shed and the affluent group least with 1.3% shed. During the low hours, the comfortable group is found to have shed 0.07% on average whereas the other two socio-economic groups have increased consumption.

Table 4-20 shows the percent treatment effect with imputed values. It shows all socio-economic groups responded to the treatment and shed consumption during the high hours; the comfortable group most with 9.04% shed and the adversity group least with 7.31% shed. All groups similarly shed during the normal hours; the adversity group most with 5.27% shed and the comfortable group least with 3.76% shed. During the low hours, the affluent group is found to have shed 0.68% on average whereas the other two socio-economic groups have increased consumption.

A similar analysis using data from January and February of 2014 can be done as shown in 4.3. The results from the analysis are in tables 4-21 and 4-22. It's important to refer back to figure 4-11; the per socio-economic group mapping holds a higher error, particularly the adversity group. Additionally, figure 4-11 is with no imputed data.

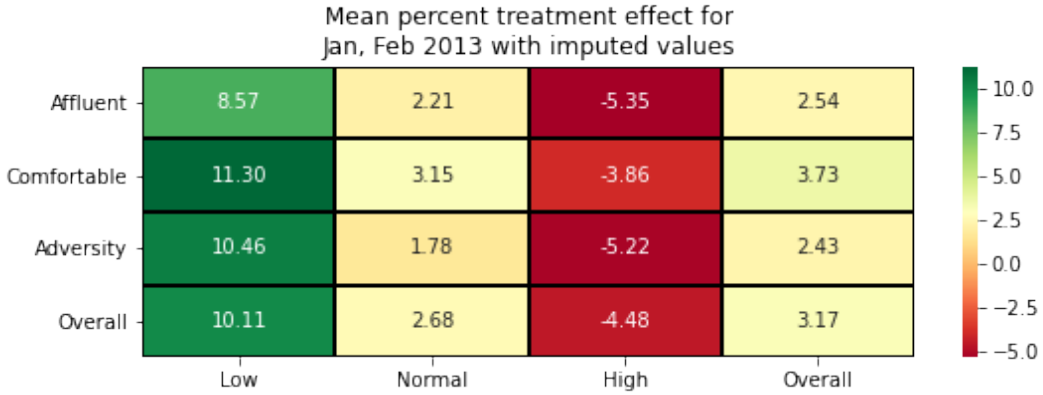

Figure 4-21: Mean percent treatment effect calculated using first two months of 2013 and 2013. Mapping is done over all houses.

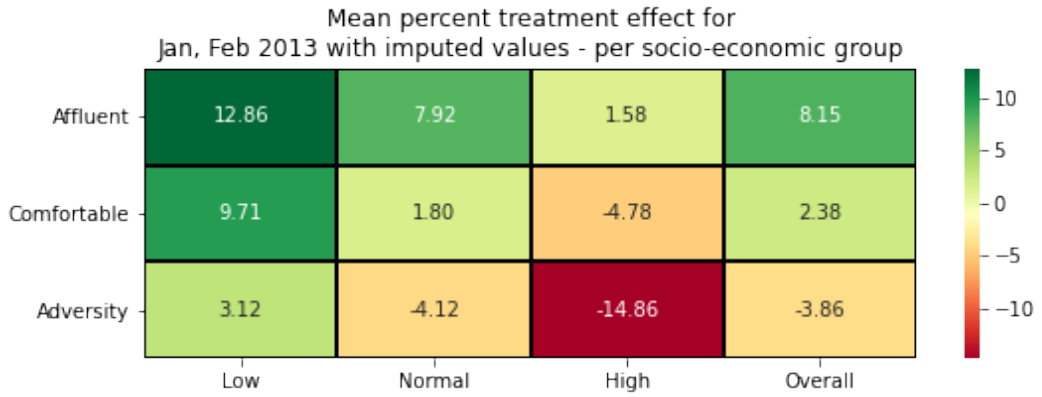

Figure 4-22: Mean percent treatment effect calculated using first two months of 2013 and 2013. Mapping is done per socio-economic group.

With the sole exception of the mean percent treatment effect for the affluent subgroup as estimated by January and February of 2014 with imputed values (table 4-22), all mean treatment effect and mean percent treatment effects show that consumption was lowered in the high price hours. This proves the original hypothesis put forth in section 3.2.

4.3.4 Limitations and Conclusion

Given the limitations of the data set as outlined in section 3.3.1 and that there is missing data for some fraction of 2012, it is difficult to strike a balance between some of the following: having a full year's worth of data to capture the temporal trend shifts within a year, removing households with missing data or imputing said missing data, and having sufficient households present in the analysis. Imputing the missing data introduces some error, and the method itself has some error which is dependent on the type of mapping (overall or per socio-economic group) as well as the size of the consumption matrices. As a result, it is difficult to confidently isolate the value

of the treatment effect. It is, however, possible to compare values to understand the relative response to the treatment in different hours and socio-economic groups. That said, every model in this chapter has a non-zero error. In the worst case, the treatment effects reported should be assumed to be $\pm$ the errors I mentioned above. At best, it's possible that those errors cancel out. Different mappings (a single mapping for all households or a different mapping per socio-economic group), whether there was any imputation, lead to different results. However, all results have consistently shown the treatment was effective; either consumption was shed in the high hours or it was increased by the least amount relative to other hours.

4.4 Aggregated Multi Linear Regression Model

Comparing the last two models described in sections 4.2 and 4.3, one aggregates over all the houses and finds a linear mapping and the other includes all the household data points before finding a mapping. There is a middle ground between the two where the house indices in the α and β matrices are divided into groups and aggregated within groups. In other words, the α and β matrices are of size $t \times n_{\text{cluster}}$ where n_{cluster} is the number of household clusters. The values in each cluster are the average consumption over the houses present in that cluster. This still results in a multi linear regression but this time, the α and β matrices are filled with time series that are interpretable, unlike in the multi linear regression model.

The α and β matrices have the following dimensions in each of the models: In the aggregated linear regression model, the original matrices are reduced down to vectors of size t. Each value is the aggregated consumption value over all households. In the multi linear regression model, the matrices are of size $t \times n$ where n is the number of houses present in the consumption matrix. In the aggregated multi linear

regression model, the matrices are of size $t \times n_{\text{cluster}}$.

4.4.1 Error Analysis

Similar to section 4.3.2, to find the error on this model, I find X (as shown in equation 4.4) on the training set and find the mean percent error on the test set. To transform the consumption matrices, I take the mean over every n index. To assure that the matrix dimensions work out when doing the mapping per socio-economic status, I separated the matrices into different socio-economic groups and then took the mean over every n index.

With the exception of the comfortable group, the error for all groups has decreased. The difference between the error on the comfortable group at 0.7 training size is less than 1%. These numbers show that the aggregated multi linear regression model is more accurate than the multi linear regression model which was itself more accurate than the aggregated linear regression model. One question that may come up is how the granularity of aggregation effects the test error. Figure 4-24 shows the mean absolute percent error on every 10, 50, and 100 houses in the consumption matrices being aggregated as well as the error from the multi linear regression model added. The reason I use mean absolute percent error (instead of mean percent error as I have thus far) is to make it easier to compare percent error across the different runs of the model as a negative or positive percent error here is not of interest and only the absolute value matters. On average, the $n = 50$ performs best, second to that is $n = 100$. Every run of the aggregated multi linear regression performs better than the multi linear regression. It is important to note that at $n = 100$ the size of the matrices are quite small: $(5, 2784)$, $(6, 2784)$, and $(7, 2784)$ for different socio-economic groups and since the train and test sets are divided over index, this

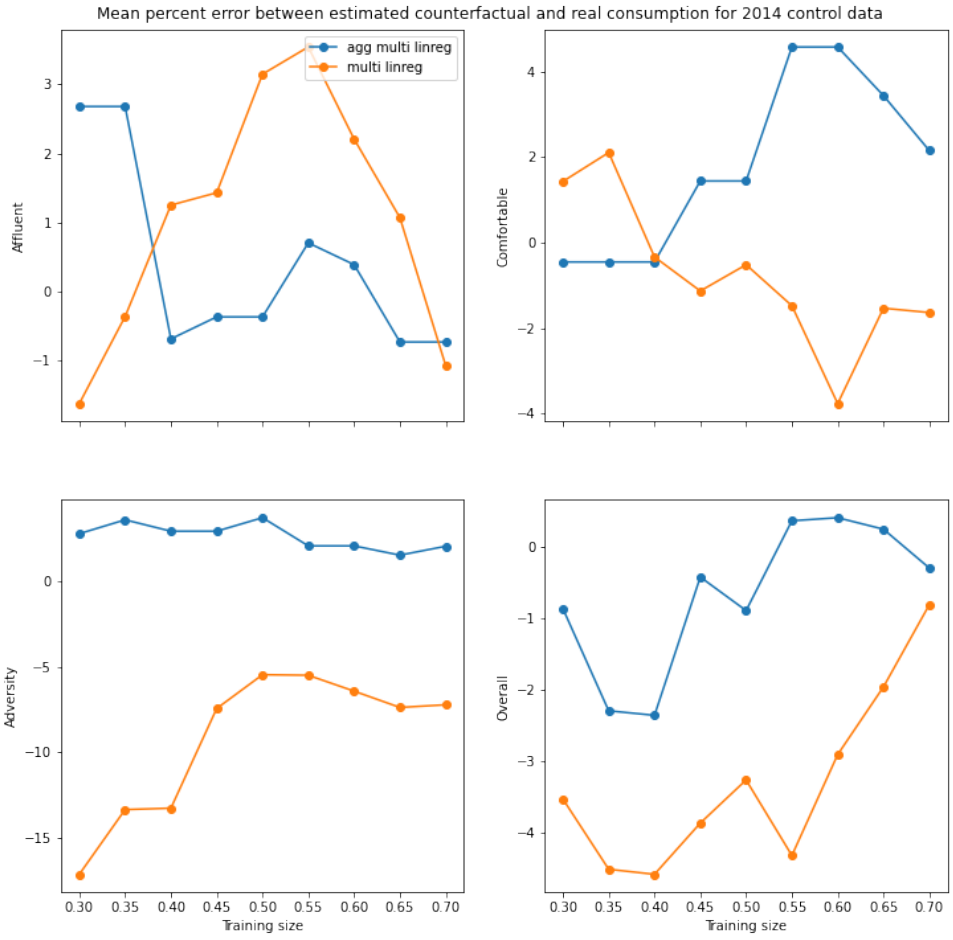

Figure 4-23: The mean percent error between the real consumption values and estimated counterfactual consumption on the 2014 test set. This analysis is done for every 50 indices in the α and β matrix aggregated to create one row. The mapping is found per socio-economic group and for the entirety of the data. Overlaid is the error for the same data without any aggregation as shown in figure 4-11.

is approaching the maximum number of indices over which we can aggregate.

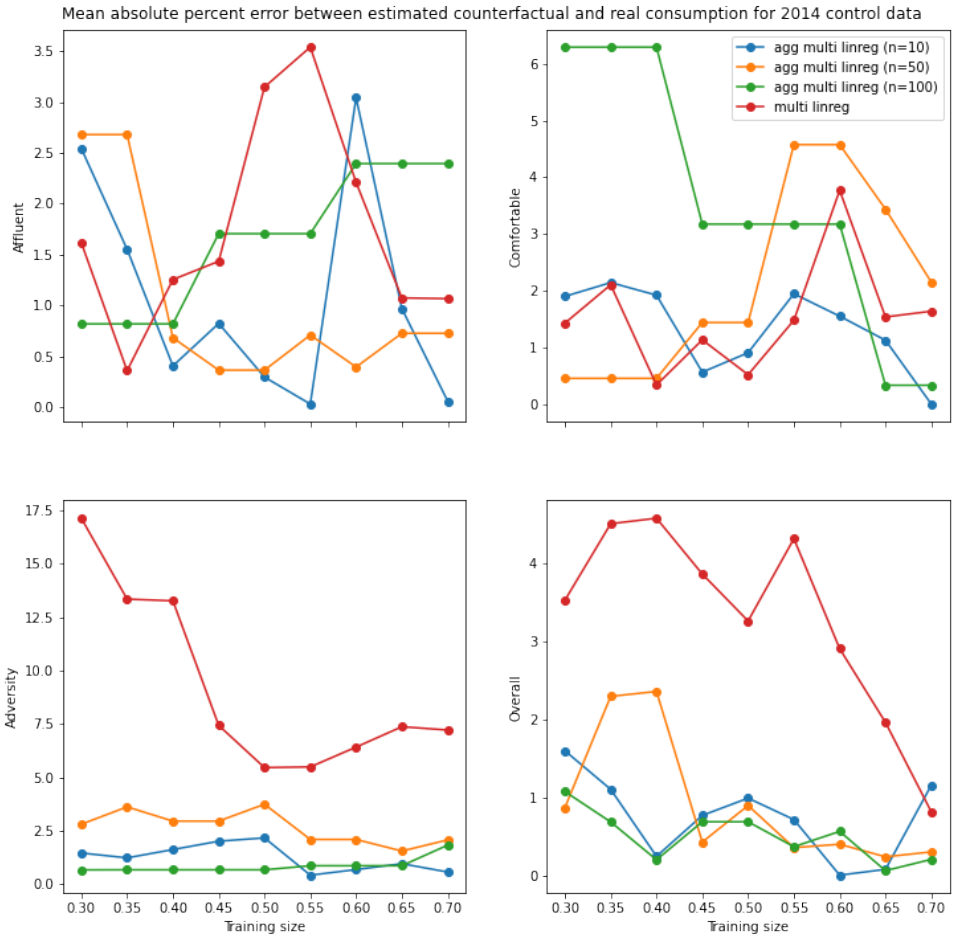

Figure 4-24: The absolute mean percent error between the real consumption values and estimated counterfactual consumption on the 2014 test set. This analysis is done for every 10, 50, and 100 indices in the α and β matrix aggregated to create one row. The mapping is found per socio-economic group and for the entirety of the data. Overlaid is the absolute mean percent error for the same data without any aggregation as shown in figure 4-11.

4.4.2 Counterfactual Analysis

Tables 4-25 and 4-26 shows the mean error and mean percent error for the 2013 treatment group. These values were found by taking the mean per socio-economic group of the dropped α and β matrices and finding a single mapping. The aggregation is done in such a way that the α and β matrices have only a single row per socio-economic group. Tables 4-27 and 4-28 show the mean error and mean percent error for the matrices with imputed values before taking the mean per socio-economic group and finding a single mapping. The results here prove the hypothesis that the treatment was effective in lowering consumption in the high price hours regardless of the mapping or the socio-economic status.

Comparing these values with similar ones in section 4.3 (tables 4-15 vs 4-25, 4-16 vs 4-26, 4-15 vs 4-27, 4-16 vs 4-28), the patterns are similar. To be more specific, the percent treatment effect resulting from the matrices with dropped values both show most shed by the comfortable group and least by the affluent group in the high hours. Similarly, both show highest percentage increased by the affluent group in the low hours and most shed in the comfortable group. Comparing the percent treatment effect resulting from the matrices with imputed values, both show most shed by the affluent group in the high hours and least shed by the adversity group, though the spread is wider in the aggregated multi linear regression method. In the low hours, the analysis on the imputed matrices both show most increase by the comfortable group and most shed by the affluent group in the low hours. In summary, the patterns of relative behavior of the different socio-economic groups throughout the different hours is the same using the multi linear regression and the aggregated multi linear regression.

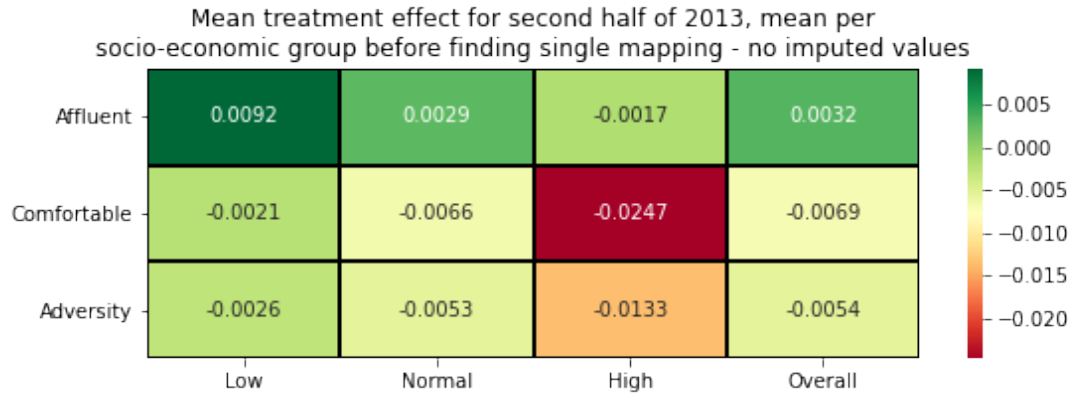

Figure 4-25: Mean treatment effect on the second half of 2013, a single mapping applied to aggregated consumption matrices per socio-economic group. The matrix upon which the aggregations were done has no imputed values.

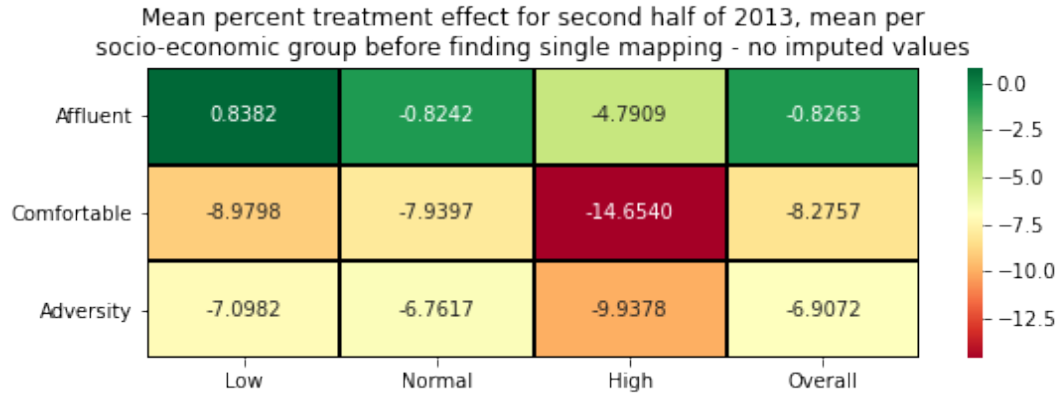

Figure 4-26: Mean percent treatment effect on the second half of 2013, a single mapping applied to aggregated consumption matrices per socio-economic group. The matrix upon which the aggregations were done has no imputed values.

4.4.3 Limitations and Conclusion

As described in chapter 1 and section 3.3, synthetic control is a method used to find the treatment effect when there are biases in the treatment and control samples or if there is no explicit control group. Additionally, synthetic control is usually applied

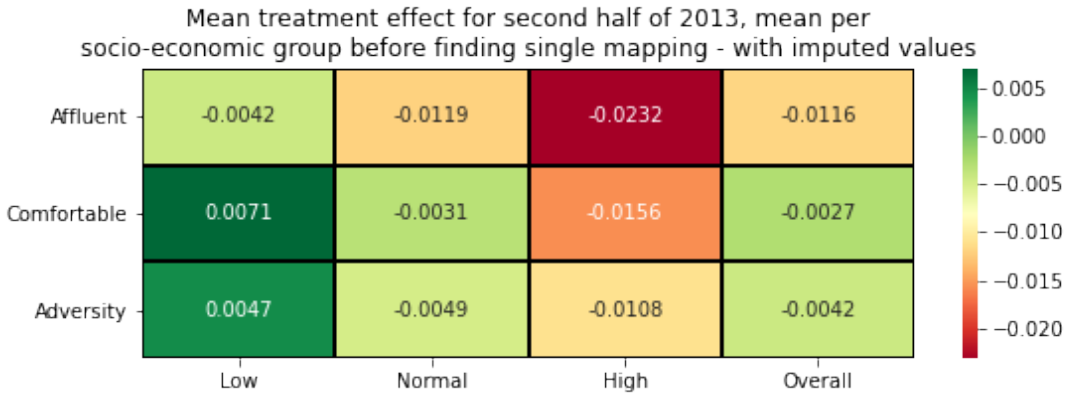

Figure 4-27: Mean treatment effect on the second half of 2013, a single mapping applied to aggregated consumption matrices per socio-economic group. The missing values in the consumption matrix were imputed before aggregation per socio-economic group.

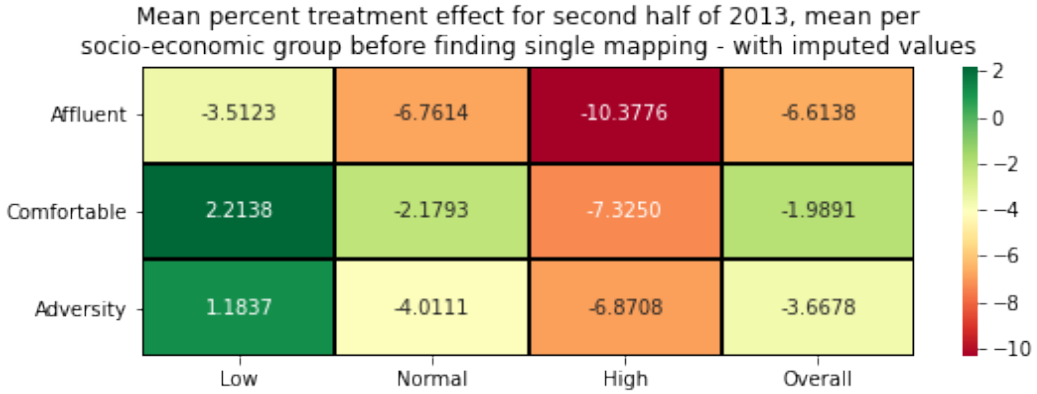

Figure 4-28: Mean percent treatment effect on the second half of 2013, a single mapping applied to aggregated consumption matrices per socio-economic group. The missing values in the consumption matrix were imputed before aggregation per socio-economic group.

in aggregate and to study an aggregate effect. In that way, the aggregated multi linear regression method is applying synthetic control as designed.

In the past three models (aggregated linear regression, multi linear regression,

and aggregated multi linear regression), we have been playing with two main hyper parameters or levers: one is whether the mapping is applied to all houses and later separated into different socio economic groups or a different mapping per socio-economic group. The other is the granularity level i.e. how many data points to aggregate over. On one end of the spectrum, the aggregated linear regression model aggregates over all households, on the other end, the multi linear regression model keeps all values. As errors show (see figures 4-23 and 4-24), somewhere in the middle is optimal and has least model error.

4.5 Next Steps: Implementing a Constrained Optimization Model

Though the multi linear regression model resulted in counterfactual values per half-hour per house, I concluded in section 4.3.2 that the results make sense when looked at aggregated over a subset of houses. However, it's a logical extension that the more accurate each counterfactual value is, the more accurate the aggregate would be. In an exploration of the $\hat{\beta}_{2013}$ from the previous model, I learned that there exist some negative values. Figure 4-29 shows the distribution of the predicted values.

This result suggests that the counterfactual consumption of some houses in some time indices would have been negative. This interpretation isn't correct as the model was never intended to be predicting the counterfactual consumption on a house level; it is meant to estimate the counterfactual consumption in aggregate. However, in order to prevent these negative values, I can solve a constrained optimization.

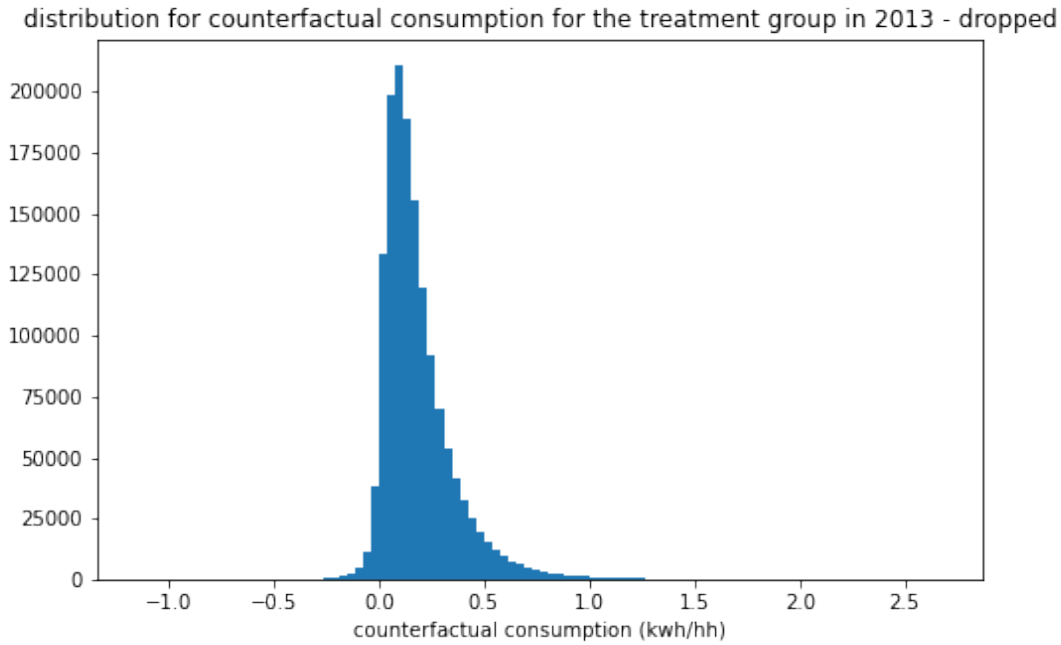

Figure 4-29: The distribution of the estimated counterfactual consumption values for the treatment group in 2013, per half-hour and per house. There exist some negative estimated values.

$$\begin{aligned} \min_{X} \quad & \|\alpha_{2012} X - \beta_{2012}\| \\ \text{s.t.} \quad & 0 \leq \alpha_{2013} X \leq max \end{aligned} \tag{4.5}$$

We can find the counterfactual consumption in 2013 $\hat{\beta}_{2013}$ using the X found from above. Figure 4-30 shows $\hat{\beta}_{2013}$ aggregated over a subset of the houses found through constrained optimization as well as without the constraint. As discussed in section 4.3.2, the results only make sense when aggregated.

Given the size of the consumption matrices ($n\times \sim 10,000$), this optimization

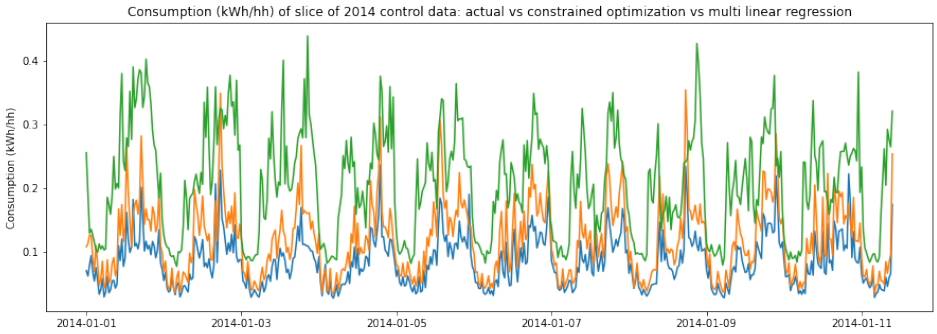

Figure 4-30: Results of the constrained optimization vs the multi linear regression method vs the actual values. This is for a slice of size 10×500 of the 2014 control data.

takes a lot of computational power. An interesting future line of inquiry is running the above optimization for the entire consumption matrices with and without imputed data and comparing the resulting treatment effects.

4.6 Conclusion

The regression models in this chapter explore the effects of granularity of data on the accuracy of treatment effects. At the one end, the aggregated linear regression aggregates over all the data. On the other end, the multi linear regression model keeps and uses all data points. Table 4.3 compares the error between the models. All the models were applied both to find a single mapping for all houses and a mapping per socio-economic status: the error for both mapping is included in table 4.3. The aggregated multi linear regression performed best, second was the multi linear regression model, and last was the aggregated linear regression model.

The models predict a counterfactual consumption for the treatment group in 2013. Comparing this value to their actual consumption in 2013, I find a mean percent

Error Analysis on the Control Group in Mean Percent Error (MPE)		
Model	MPE	Train size
Aggregated linear regression (overall)	-2.58%	0.7
Aggregated linear (affluent)	-2.68%	0.7
Aggregated linear (comfortable)	-3.92%	0.7
Aggregated linear (adversity)	-3.38%	0.7
Multi linear regression (overall)	-0.81%	0.7
Multi linear (affluent)	-1.07%	0.7
Multi linear (comfortable)	-1.64%	0.7
Multi linear (adversity)	-7.21%	0.7
Aggregated multi linear regression (overall)	-0.30%	0.7
Aggregated multi linear (affluent)	-0.73%	0.7
Aggregated multi linear (comfortable)	2.15%	0.7
Aggregated multi linear (adversity)	2.06%	0.7

Table 4.3: Error analysis for different regression models used on the control group reported in mean percent error (MPE).

treatment effect. The reason I chose that metric is because it factors the group's actual consumption in — what percentage of their consumption did they change as a response to this treatment is a much more meaningful metric vs a metric in kWh/hh. In this section, given that there was missing data, an imputation function was used to fill the missing values. All models conclude that the treatment was effective: all socio-economic groups either shed consumption in the high hours or increased consumption by the least amount, relative to other hours. The models are not all in agreement in terms of which socio-economic group is the most price sensitive (shed consumption most in the high hours). However, eliminating other manipulations that could introduce error (missing value imputation) a consistent outcome is that the comfortable group shed demand most (refer to figures 4-14 and 4-18).

Chapter 5 outlines a review of time series models and the details of a random forest regression (RFR) model that learns the temporal trends as well as dependence

on temperature. Lastly, I will be comparing the results from all the models in this chapter as well as the RFR model in chapter 6.

Chapter 5

Models & Results: Time Series Prediction Models

Going back to chapter 1, demand response models and literature either estimates baseline consumption or predicts future load. Chapter 4 outlined four regression models that found mappings to find the counterfactual consumption of the treatment group in 2013. Those models do not have an explicit time dependence. Time series models that incorporate an explicit temporal dependence and can take other regressors (such as temperature) theoretically may have better predicting powers since consumption is both time dependent and temperature dependent.

The ability to predict consumption is critical. It allows system operators to know the severity of upcoming peak hours and offer well-timed, well-priced, and targeted DR incentives. Additionally, predicting consumption can be used to estimate the counterfactual — what would have been consumed without the treatment — to judge if a treatment was effective.

In this chapter, I outline a review of different time series models that I tried

for this data set, as well as the most promising results. I close the chapter with a comparison between regression models and time series models.

5.1 Model Review

In this section, I will review and outline a number of time series prediction models that I attempted. I chose a random forest regression model for the analysis that I will go into in detail in section 5.2.

5.1.1 Identifying Features: A First Principles Approach

An energy model derived from first principles depends on the following active and passive consumption [32]. In other words it depends on the following features:

- time of day
- day of the week
- day of the month
- month of the year, which is synonymous with seasons (and correlated with temperature)
- temperature
- price per kWh of consumption
- consumer specific features such as socio-economic status, number of people in a household, etc.

Given all these dependencies, one approach is to run a machine learning model to find the hourly consumption per house. Another option is to fit a multi-dimensional

higher order polynomial to the consumption data. I will explore a very simple approach to a consumption model below.

In order to see if the consumption takes affect by the above features, I used the data from the trial. All the above affected level of consumption except for day of the month. Going back to the first principles approach, there's no reason why the day of the month would have an effect on the consumption. Temperature has an inverse relationship with consumption, as shown in figure 5-10, in this data set because heat in London is electric.

Temperature wasn't part of the original data set but I was able to get hourly temperature data from the Power Data Access Viewer by NASA [24]. To do half-hour analysis, I've extrapolated the same hourly temperature to be present throughout that hour. I assumed the same temperature for all houses and the temperature in London as found on the Power data base.

The following model is a simplification and separates the dependence of time and temperature.

$$f(T,t) = g(t) + h(T) + \epsilon$$

where $g(t)$ is the time dependence and a fitted sum of sine functions and $h(T)$ is the temperature dependence on the remainder another fitted function and an inverse relationship as previously stated. Figures 5-1 and 5-2 show a sum of sine fitted to the temporal component — $g(t)$. Figures 5-3 and 5-4 show a $\frac{1}{T}$ and linear fit to the temperature component — $h(T)$. Figure 5-5 tries to predict the residuals. The above method assumes independence between the temporal and temperature components which is a significant simplification. Additionally, in order to perform well, this functional fit would need to be applied to different segments of the data (socio-economic groups, seasons, etc.). I explore some other time series prediction

models below.

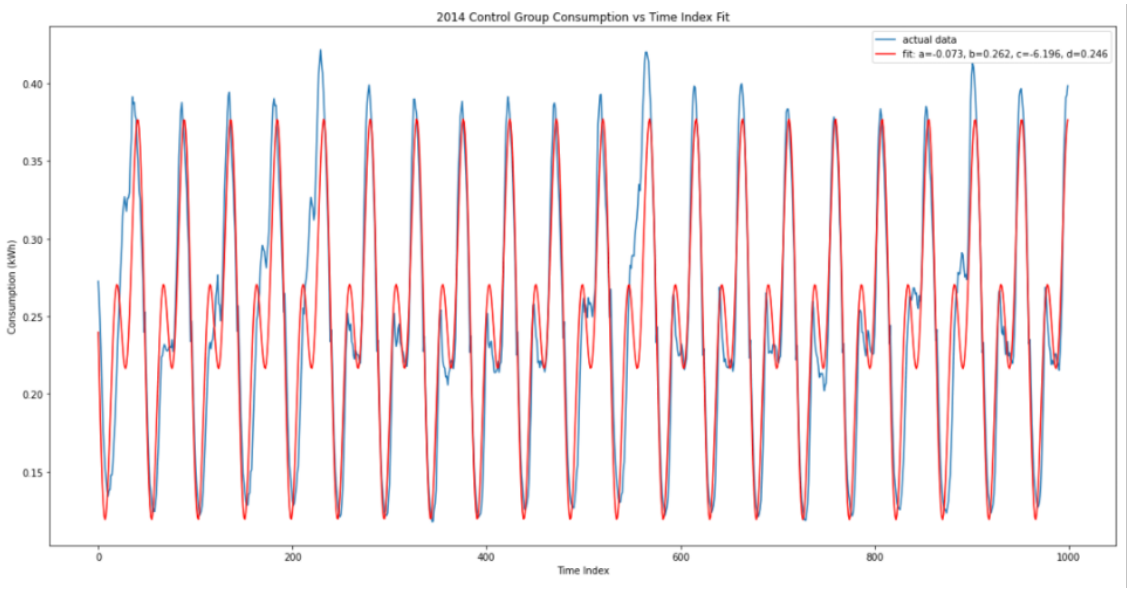

Figure 5-1: Sum of sines fit to the half-hour level consumption data, 2014 control group.

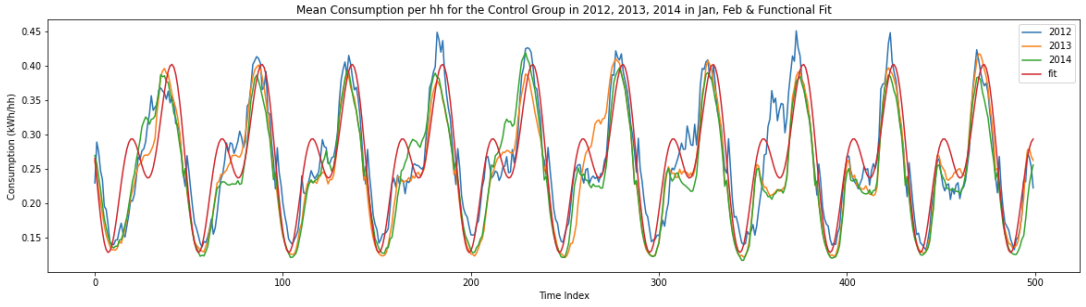

Figure 5-2: Sum of sines fit to the half-hour level consumption data overlaid by control data from Jan, Feb 2012, 2013, 2014.

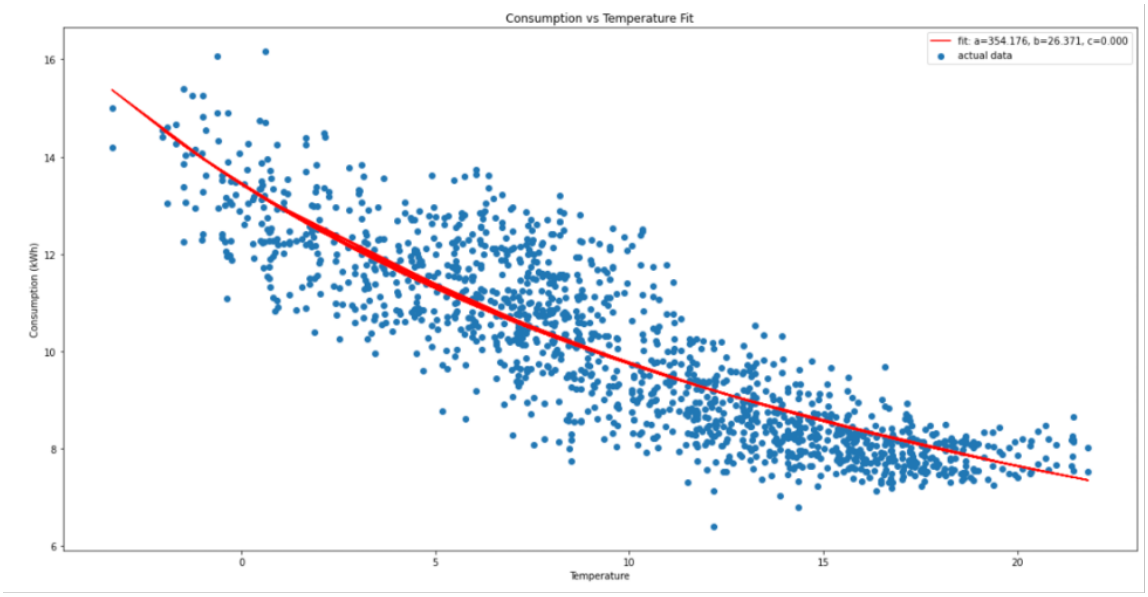

Figure 5-3: 1/T fit to the residuals.

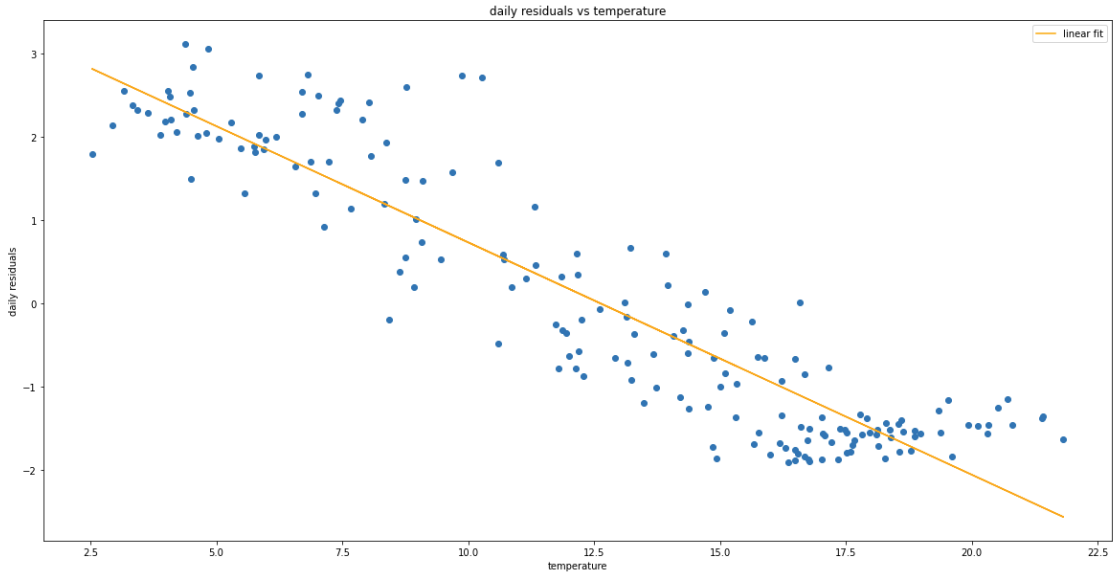

Figure 5-4: Linear fit to the residuals.

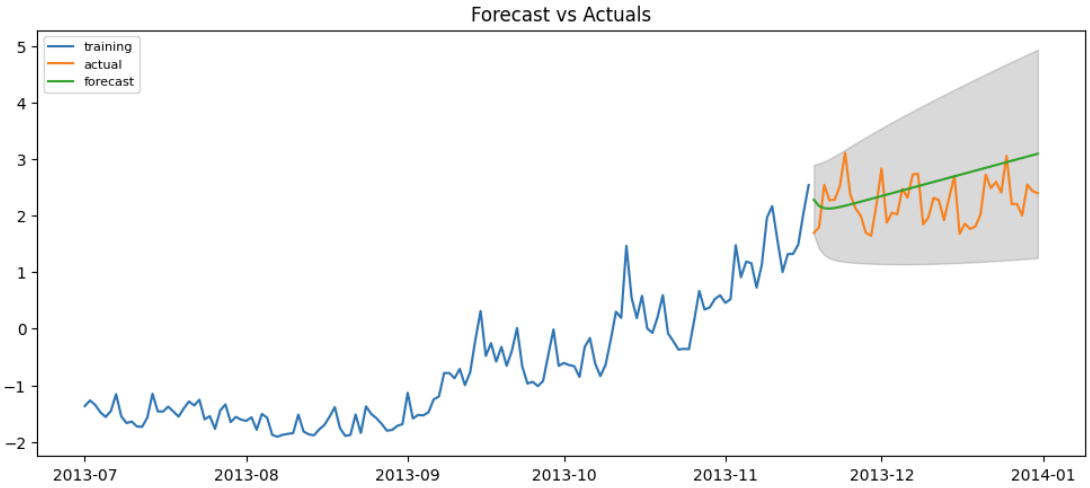

Figure 5-5: ARIMA fit on the residuals.

5.1.2 Autoregressive Integrated Moving Average (ARIMA) Model

A popular and widely used statistical method for time series forecasting is the ARIMA model. The ARIMA models combines correlation between lagged observations (AR), constant difference between raw observations (I), and univartiate linear dependence on past observations (MA). Given the seasonal behavior of electricity consumption, an ARIMA model is a good choice to construct the time dependent components of the consumption model. There are multiple levels of time dependence present in our data set:

- time of day (as shown in figure 5-11)
- day of week (as shown in figure 5-11)
- month of the year (as shown in figure 5-10)

I trained an ARIMA model on the 2012 control group's average hourly consumption and tested on the same metric in 2013. Additionally, hourly temperature from

the power API was included as an exogenous variable. Temperature was deemed to be a significant variable because the London heating system is electric, causing high demand for energy in winter months and relatively low demand in the summer. The goal is to be able to train a model on the 2012 control data that has low error on the 2013 control data. If this is doable, then I can similarly train a model on the 2012 treatment data and assume that with high accuracy, the output will be the 2013 counterfactual consumption.

A non-seasonal ARIMA model is classified as an ARIMA(p,d,q) model, where: p is the number of autoregressive terms, d is the number of nonseasonal differences needed for stationarity, and q is the number of lagged forecast errors in the prediction equation. For finding p, if the ACF goes into negative, then we can conclude that the time series has been over-differenced. In ARIMA models, we're basically regressing on lags, hence it works best if said lags are independent.

An ARIMA model with order (2,1,1) was chosen using the `auto-ARIMA` Python package, accounting for the fact that the data is highly seasonal. The model predicted 2013 data poorly without the exogenous variable of temperature; the Mean Absolute Percentage Error (MAPE) was 23%. By adding the exogenous variable of average daily temperature, the model performed better but was still very inaccurate with a MAPE of 19%. Given the previous sections on the missing data in 2012, perhaps the model can perform better if trained on more populated data and tested on a smaller time period.

5.1.3 Prophet: Automatic Forecasting Procedure

The Prophet forecasting model is a time series forecasting model that is open source and released by Facebook's Core Data Science team [33]. The model is a modular

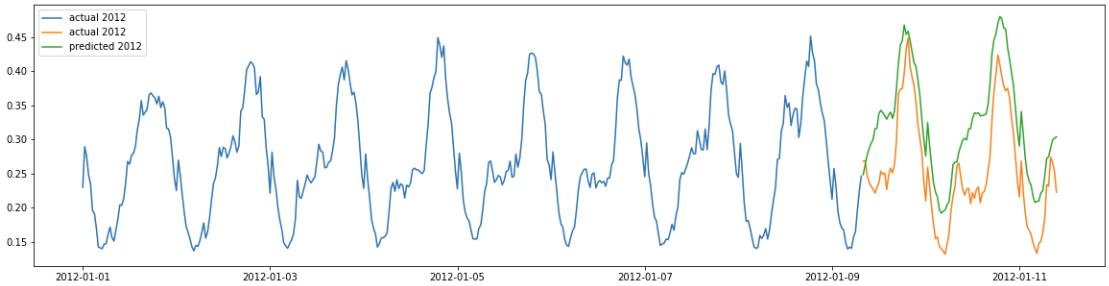

Figure 5-6: The model has been trained on the control data from the first 9 months of 2012 and tested on the remainder. The model cannot be extended to the 2013 control data and performs very poorly.

additive regression model with interpretable parameters where non-linear trends are fit with yearly, weekly, and daily seasonality. It works best with time series that have strong seasonal effects and several seasons of historical data. It has the ability to handle shifts in the trend, outliers, and is robust to missing data. This seems like a perfect model for our data. It can take care of different levels of seasonality, take on regressors such as temperature, and incorporate holidays. The GitHub for the model (linked) was helpful in setting up the model for this use case.

Figure 5-7 shows the model training on part of the 2012 control data and its performance on the remainder. The error on the test set was -3.77%. Figures 5-8 and 5-9 shows the components of the model found in figure 5-7. The daily component that has been found is very similar to patterns observed in this data set as seen in figure 5-11.

The upwards trend in the 2012 data (as seen in figure 5-7) results in the 2013 data continuing on that upward trend. As a result, it performs very poorly on the 2013 control data. This is perhaps due to the fact that Prophet requires historic data, specifically to pull out annual and temperature trends, and given the missing data, we can only train the model on the latter half of 2012.

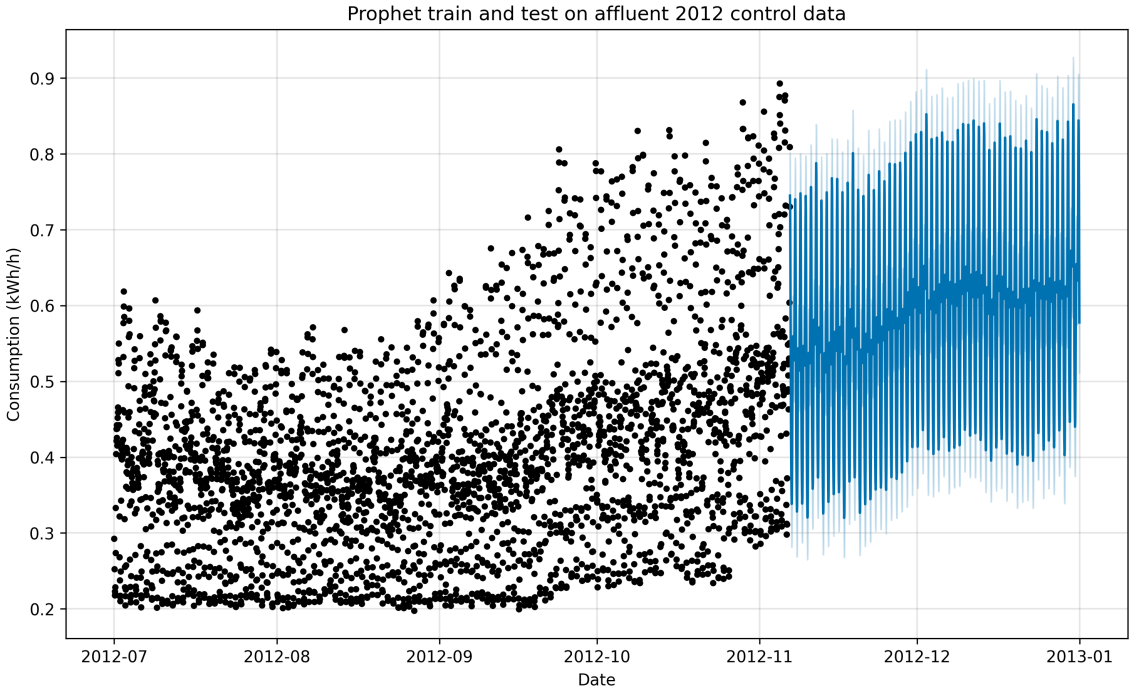

Figure 5-7: Prophet prediction on the affluent 2012 control data, trained on 70% and tested on the remainder (56 days).

5.2 Random Forest Regression (RFR) Model

Models outlined in chapter 4 have no explicit temporal trend and make no explicit use of temperature data. Another way to estimate the counterfactual consumption is to learn a time series model based on the data in 2012 that also incorporates temperature data. Temperature wasn't part of the original data set but I was able to find hourly temperature data for London for the entire duration of the trial from the Power Data Access Viewer by NASA [24]. It is assumed that the temperature is known to help more robustly estimate average consumption in 2013. Temperature is a significant variable because the London heating system is electric, causing high demand for energy in winter months and relatively low demand in the summer as

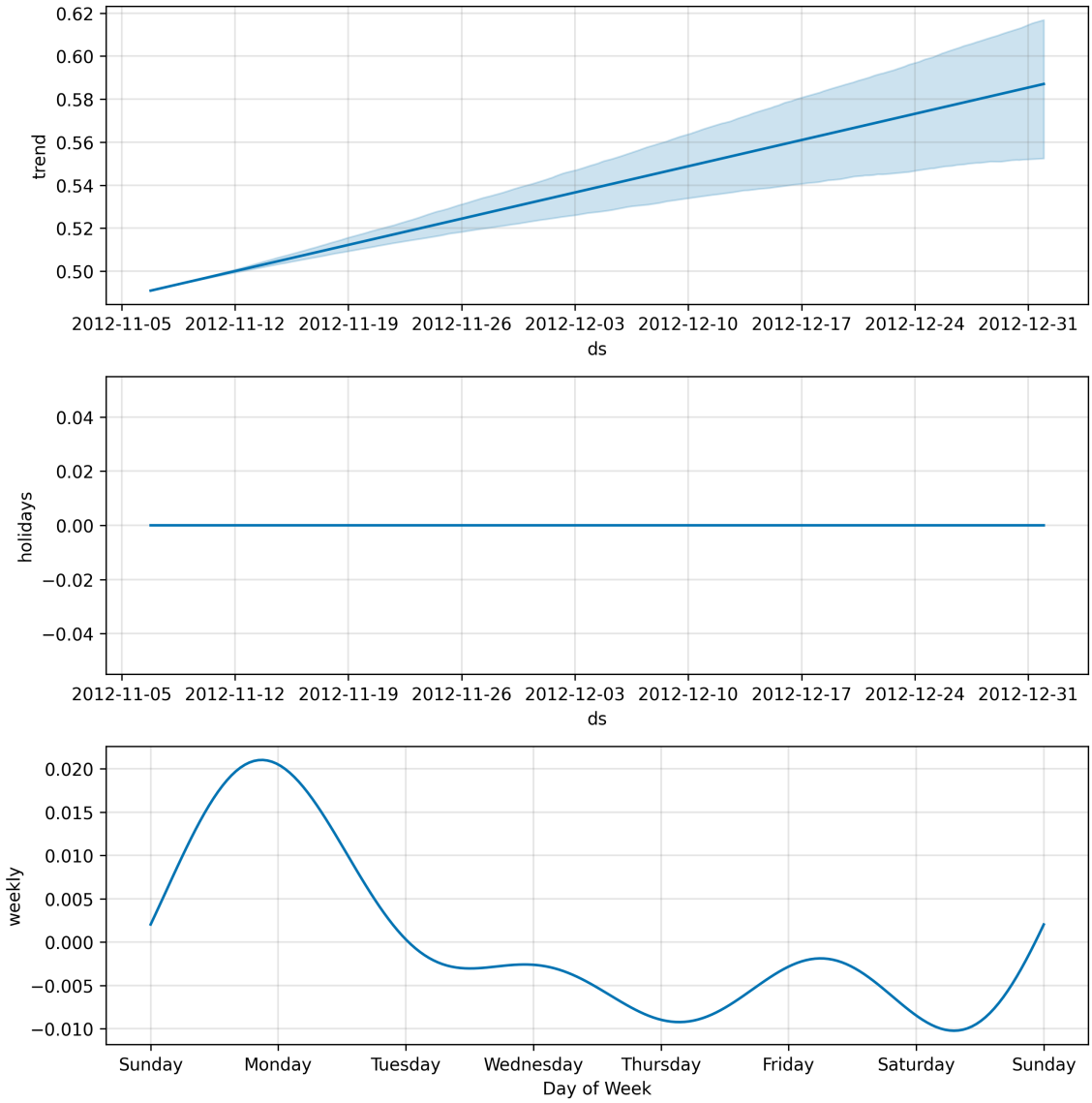

Figure 5-8: The trend, holidays, and weekly components of the prophet model seen in 5-7.

seen in figure 5-10.

As figure 5-11 shows, consumption looks differently on weekends and weekdays. However, the daily consumption wave form looks similar throughout different seasons.

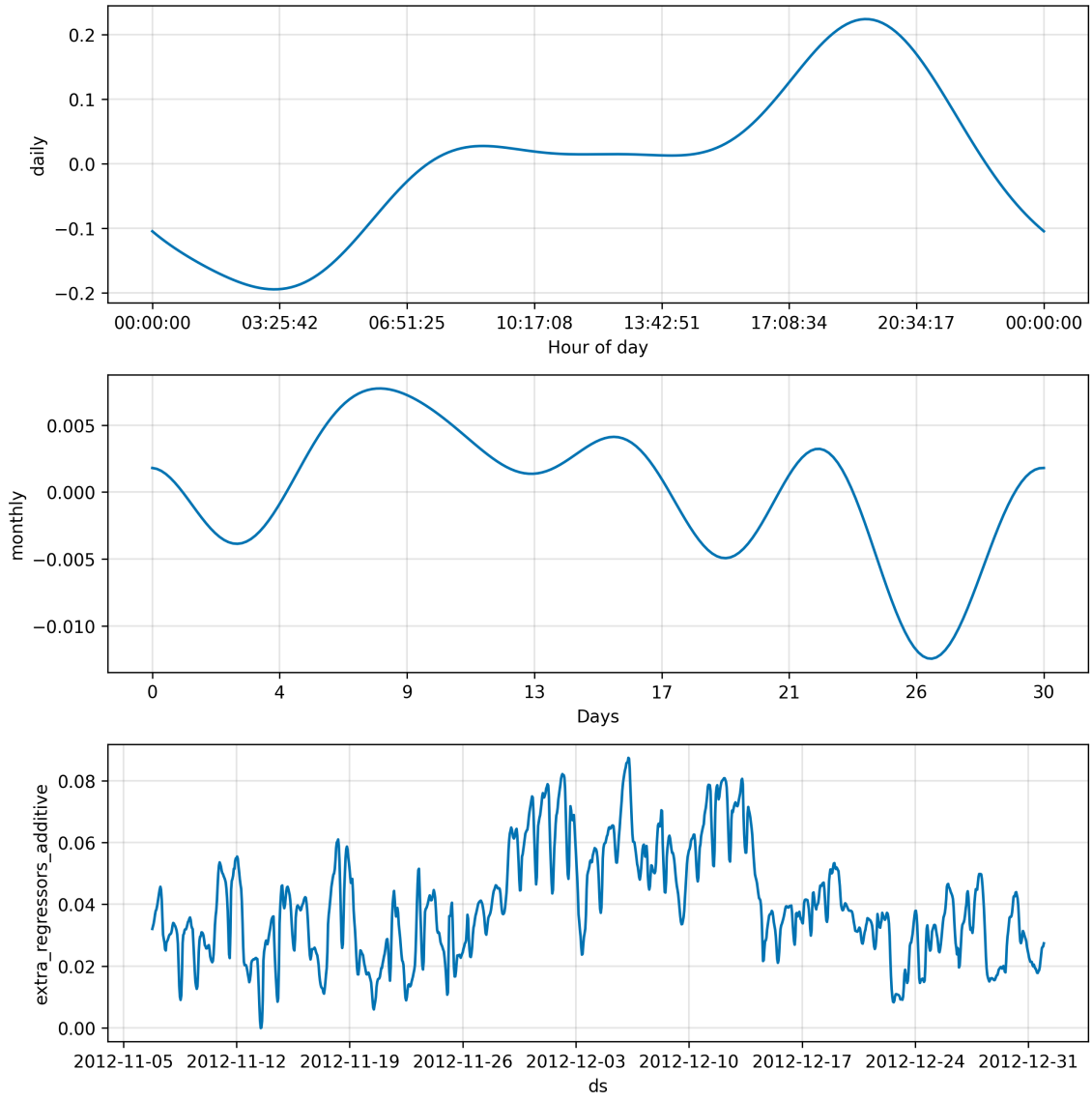

Figure 5-9: The daily, monthly, and temperature components of the prophet model seen in 5-7.

This is verified by finding the singular values of the matrix that holds 24 hour level data for all days in the second half of 2012 as can be seen in figure 5-12. Most of the data is explained by the first principle component. I will refer to this matrix as the

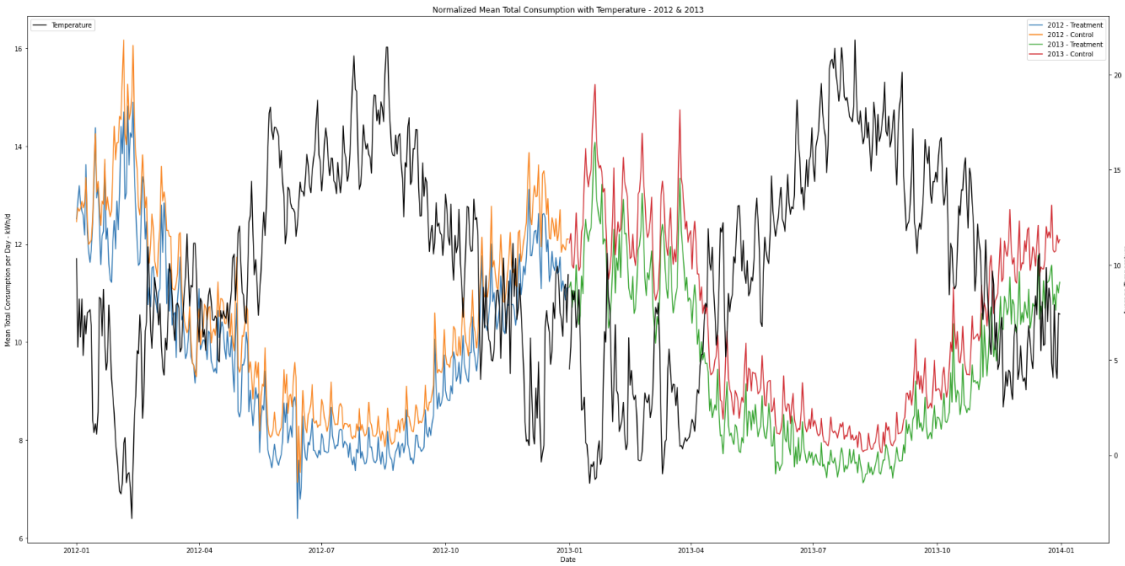

Figure 5-10: Average consumption per day with temperature. This shows an inverse relationship between consumption and temperature since London uses a lot of electric heating; consumption is higher in colder seasons and lower in hotter seasons.

consumption matrix going forward. It is of dimension number of days in the subset of the data $\times$ 24.

So far, a dependence on month of the year, temperature, day of the week, and socio-economic status can be observed. I am using a random forest regressor to learn the temporal trends and include the named features in the training data. A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the data set and uses averaging to improve the predictive accuracy and control over-fitting.

I train a random forest model per socio-economic group and with month of the year, temperature, day of the week as input samples. In order to get a daily granularity for my input data I use a minimum, maximum, and mean temperature value per day. The values in the lower dimensional representation of the consumption matrix

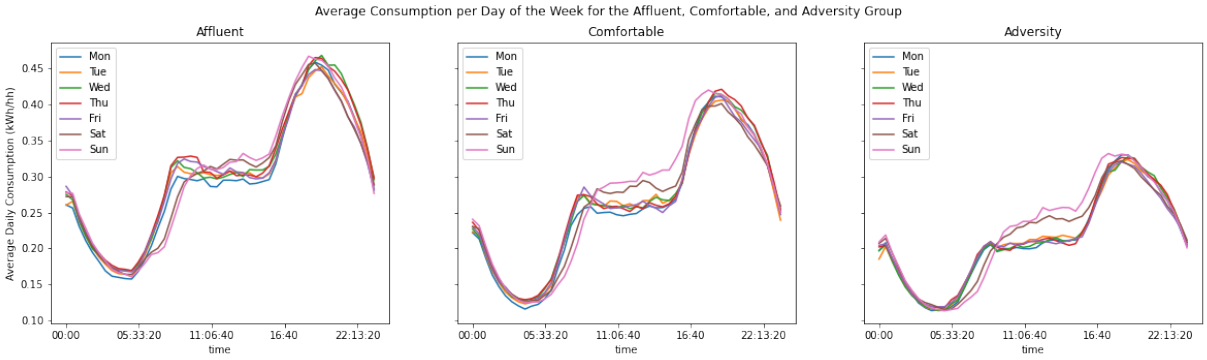

Figure 5-11: Average consumption throughout the day over weekdays vs weekends broken down by socio-economic status. As can be seen the consumption pattern looks different on weekdays and weekends and is most in the affluent group and least in the adversity group.

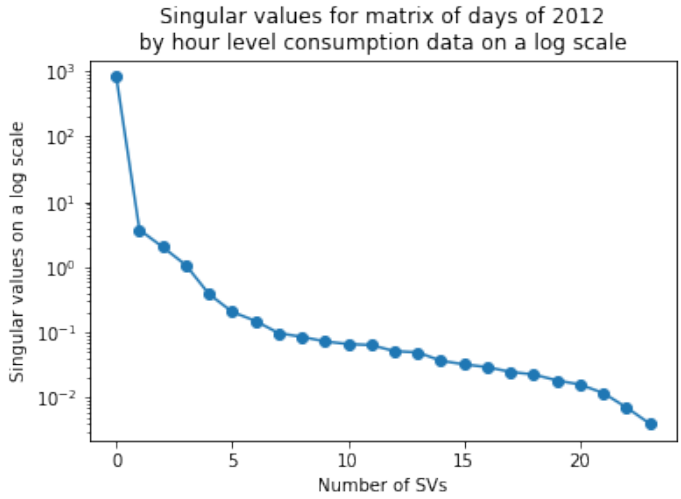

Figure 5-12: Singular values for control group's consumption matrix of days of 2012 by hour level data (the comfortable socio-economic group). The SVs of the affluent and adversity groups are similar.

are the target values.

Given that the daily trends in the consumption matrix can be captured by the first principle component, as shown in figure 5-12, I project the 2012 control group

consumption data down to that principle component. I then split both the input data and this lower dimensional representation of consumption into a training set and a test set. The trained model, therefore, now goes from the input data to a lower dimensional representation of the consumption matrix.

5.2.1 Error Analysis

All the data used for this model is for the control group as the goal is to learn a robust time series model of consumption for 2012/2013. The model trained on the comfortable group predicts on the 2012 test data with 0.29% error and on the 2013 data with 4.77% error. This result shows that this time series model can similarly predict 2013 treatment consumption within ∼5%. I will be training a similar model on the 2012 treatment data in order to find an estimate for the counterfactual consumption of the treatment group in 2013. Table 5-13 shows the test error and error on the 2013 data per socio-economic group. I trained the model on up to the first four principle components with the data for each socio-economic group to find what number of principle components minimized the error on the 2013 data. That information is also included in table 5-13 as well as the test size.

	Test error	Error on 2013	Number of PC that minimizes error on 2013	Test size
Adversity	0.46	3.63	1	0.2
Comfortable	0.29	4.77	1	0.2
Affluent	0.39	5.00	1	0.2

Figure 5-13: Test error and error on the 2013 data per socio-economic group as well as the number of principle components that minimized error on the 2013 data and the test set.

5.2.2 Counterfactual Analysis

Following the results on the control group, I ran a similar model per socio-economic group on the 2012 treatment data and found the 'error' on the 2013 treatment data — this is the treatment effect plus the error on the model; similar to other models. The models had similarly low test error as was the case when it was run on the control group (shown in figure 5-13). I used only one principle component for these models as that lead to lowest error on 2013 data on the control group and had a test size of 0 i.e. the model was trained on the 2012 treatment data, entirely. Tables 5-14 and 5-15 show the mean treatment effect and mean percent treatment effect per socio-economic group.

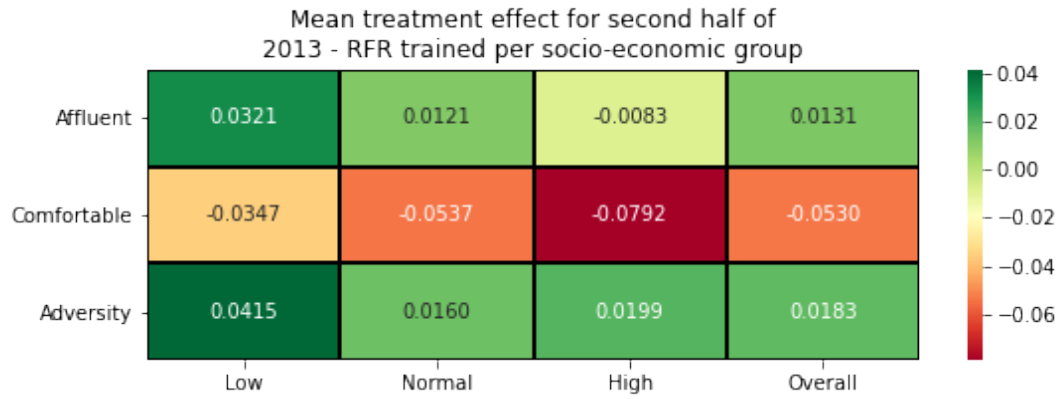

Figure 5-14: Mean treatment effect for the second half of 2013. The random forest model was trained on the second half of 2012 treatment data.

So far we've been using temperature (minimum, maximum, and mean values per day), day of week, and month of year as input data. Given figure 5-11, it might make more sense to a binary column that differentiates between weekdays and weekends. Additionally, it might be a good idea to use a reduced version of the daily control data as input values as well (the first two principle components). Going back to equation

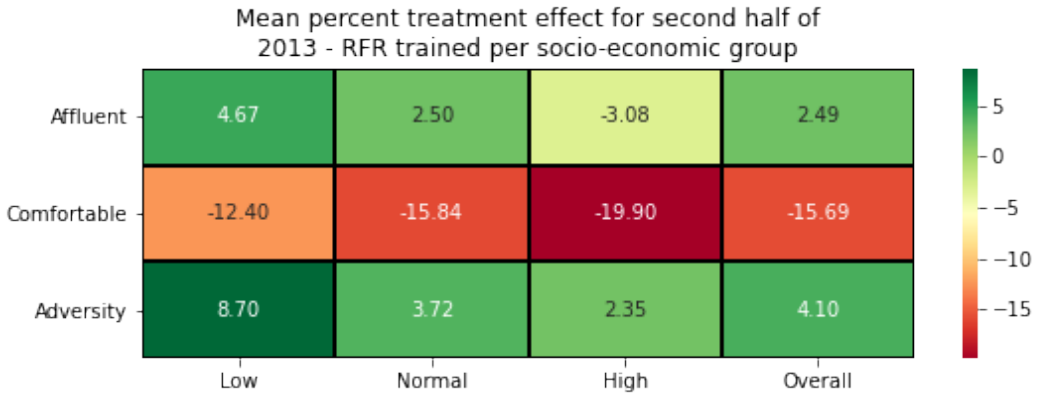

Figure 5-15: Mean percent treatment effect for the second half of 2013. The random forest model was trained on the second half of 2012 treatment data.

3.3, our goal is to learn a model on the 2012 treatment data such that once it has the 2013 input data, it can approximate the 2013 treatment group counterfactual. The idea behind using the 2012 control group data as an input to the model is that though the control and treatment groups have differences in behavior, this difference is consistent over the two years. As a result, having the 2012 control group data as input might the model better learn the 2012 treatment data. Tables 5-16 and 5-17 show results from this analysis.

The relative outcome of the two runs of the RFR model are similar. Comparing tables 5-15 and 5-17, both show shed in the comfortable group throughout the different price points. They both show that the comfortable group shed most in the high hours and the adversity group shed least. They similarly both show a decrease by the comfortable group in the low hours and the largest increase by the adversity group in the low hours.

A few questions may come up: why consider a different model per socio-economic group? Why run the model on the aggregated data per socio-economic group and

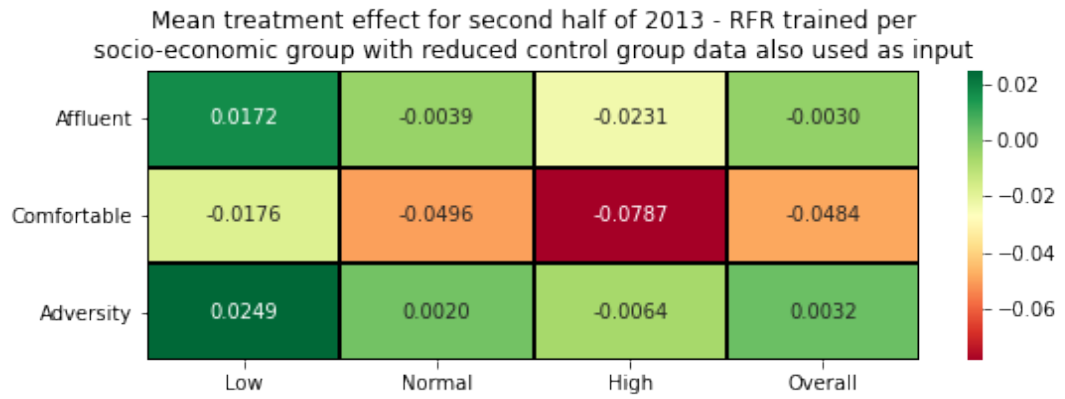

Figure 5-16: Mean treatment effect for the second half of 2013. The random forest model was trained on the second half of 2012 treatment data with the control group as input.

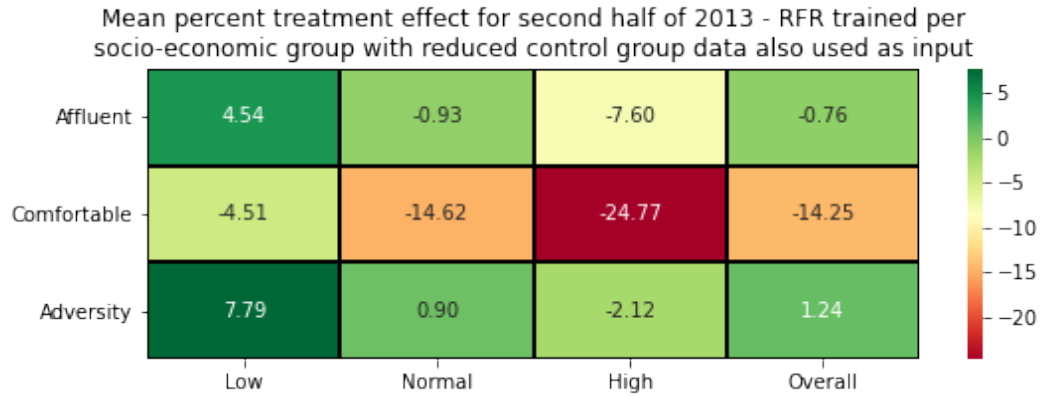

Figure 5-17: Mean percent treatment effect for the second half of 2013. The random forest model was trained on the second half of 2012 treatment data with the control group as input.

not include the house level data? Why run PCA on the consumption matrix? Below, I expand on the reasons behind my choices for this model.

The reason for training a different model per socio-economic group is three-fold. First, as shown in figure 5-18 and 5-11, consumption in different socio-economic

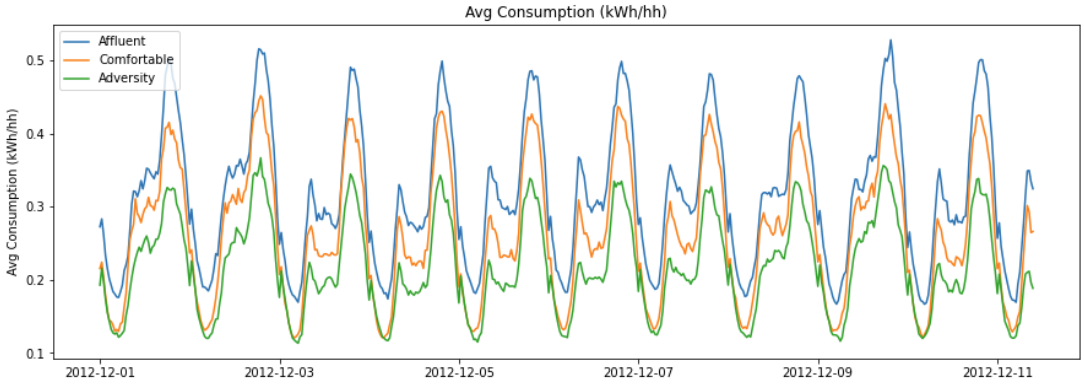

Figure 5-18: Average consumption per half-hour per socio-economic group. This figure shows the socio-economic dependence of consumption amount and consumption pattern.

groups differs significantly. Second, in the current RFR model, there is a 1-1 mapping between the input values and the output values. However, if socio-economic group is added to the input matrix, so the model could learn the dependency, the output matrix would be different per socio-economic group without a clear way to map the input matrix to the output matrix. In other words, this decision was due to technical limitations of the model. Third, intuitively, it makes sense that if the model is trained per socio-economic group, then there would be higher accuracy as it will only focus on finding a single trend. Exploring a single model that incorporates both socio-economic group and price as inputs is a potential next step for this project and could be used as a load forecasting model as well.

The reason I ran the model on the aggregated data is that the goal of this model is to capture the temporal trend present in the data and these trends only become present when the consumption data has been aggregated over some subset of households as shown in figure 4-8. Additionally, the reason behind running PCA on the consumption matrix is that as shown in figure 5-12 most of the daily consumption

trend is captured by the first four principle components. This shows that most of the data present in that matrix does not hold any information. To avoid over-fitting and add to the efficiency of the model, the data can be reduced down without any information being lost.

5.2.3 Limitations and Conclusion

This model has an explicit dependence on time and also includes temperature whereas regression models outlined in chapter 4 do not incorporate those dependency. The first iteration of the model, without the control data, can be used as a basic load forecasting model. It can predict a consumption value given temperature, day of week, and month of year. The second iteration of the model can similarly forecast load but given its dependence on the control data, it needs some historic data for the households in question. For this reason, it may be more suitable for a baseline estimation analysis and not load forecasting. Given the inclusion of the control data as an input for the second model, it's impossible to quantify the error in the model. As a reminder, this model is run on aggregate data so it can estimate the baseline or forecast the load of a group of houses, and not independently.

5.3 Next Steps: Implementing a Shift Model

The dToU treatment has resulted in both lowering consumption and shifting consumption from high price hours to normal and low price hours. A future analysis can look the dependence of this shift on socio-economic status, what time of day the high price hours were set, and seasons.

Consumption has a component that is absolutely required, another that can be

shifted around (to other times of the day or other times of the week, perhaps), and a last component that is extraneous and can be shed. Let's assume that the shift happens only throughout the day i.e. choosing what hour to wash the dishes or when to charge your electric vehicle and that there is no shift between the days. There may also be shifting going on throughout the week i.e. choosing when to wash your clothes but that is harder to capture.

The following model can be used. Let's assume that our daily consumption follows the following trend:

$$D_i = P_i + S_i + R_i + \epsilon_i$$

where

- D_i is the daily consumption
- P_i is periodic in i with period 24, accounting for hourly periodicity; this is the absolute necessary portion of consumption per day
- S_i is consumption that is needed but shift-able around the day
- R_i is consumption that is extraneous and can be reduced
- ϵ_i is the remaining residual that accounts for other influences

5.4 Conclusion

In this chapter, I reviewed a number of different time series prediction models: a first principle approach to building a time series model, an ARIMA model, and the Prophet forecasting model. I outlined the implementation details and results from a random forest regression model. The results are consistent with results found in chapter 4 in terms of relative response to the treatment, however, the RFR model

shows a more significant response to the treatment by the comfortable socio-economic group. The RFR model includes explicit dependency on the day of week, month of year, temperature, and in one iteration of the model even the control group. These dependencies are present in this data set and in electricity consumption patterns in general. Though the RFR model has a lower test error, it has lower transparency into how the branches of the tree are segmenting the data and predicting the counterfactual. In chapter 6, I will compare the accuracy of all the models outlined in chapters 4 and 5.

Chapter 6

Conclusion, Policy Implications, Limitations, and Future Work

6.1 Conclusion

In chapters 4 and 5, I outlined the four models used in this work: aggregated linear regression, multi linear regression, aggregated multi linear regression, and a random forest regression. Section 6.1.1 includes a comparison of the error and results from each.

6.1.1 Comparing Different Models: Accuracy and Results

Table 6.1 shows the different models in this thesis and the test error on each reported in mean percent error (MPE). All the error analyses are done on the control group for which we have ground truth consumption values and the model is applied to a subset of the 2012/2013 data and tested on the remainder. For the three models there are two mappings: one overall mapping which includes all houses, and another per

socio-economic group which only includes houses in that particular socio-economic group. The MPE is calculated by taking the difference between the predicted values and the actual ground truth values and dividing by the actual values to get a percent difference. This is how the treatment effects have been reported as it's a more tangible metric to show percent of electricity consumption shed vs a metric in kWh/hh space.

As can be observed, the aggregated multi linear regression model performs better than both the aggregated linear regression model and the multi linear regression model. The model includes some granularity while also predicting already aggregated values that hold meaning. The best performing model is the random forest regression which is in line with expectation as it includes both time and temperature dependence. However, the RFR model lacks transparency relative to the regression models.

Error Analysis on the Control Group in Mean Percent Error (MPE)		
Model	MPE	Train size
Aggregated linear regression (overall)	-2.58%	0.7
Aggregated linear (affluent)	-2.68%	0.7
Aggregated linear (comfortable)	-3.92%	0.7
Aggregated linear (adversity)	-3.38%	0.7
Multi linear regression (overall)	-0.81%	0.7
Multi linear (affluent)	-1.07%	0.7
Multi linear (comfortable)	-1.64%	0.7
Multi linear (adversity)	-7.21%	0.7
Aggregated multi linear regression (overall)	-0.30%	0.7
Aggregated multi linear (affluent)	-0.73%	0.7
Aggregated multi linear (comfortable)	2.15%	0.7
Aggregated multi linear (adversity)	2.06%	0.7
Random forest regression (affluent)	0.39%	0.8
Random forest (comfortable)	0.29%	0.8
Random forest (adversity)	0.46%	0.8

Table 6.1: Error analysis for different models used on the control group reported in mean percent error (MPE).

All of the models have confirmed that the treatment was effective: the percent change by the treatment is lowest in the high hours and highest in the low hours. All the comparisons are done with the same socio-economic group and for the same price point and reported as a percent difference, we can therefore eliminate any potential bias from high price hours being high consumption hours as well.

6.1.2 Broad Takeaways

In data science and statistical analysis, there are always compromises to be made. For example, in this problem, it's impossible to make use of house level data points (no aggregations) and want to include the entire year's data due to missing values. There are solutions around the limitations of the data set but each choice includes assumptions that would then propagate error into the models and analyses.

Methods of analysis will often have an error associated with them. It's important to have a way to estimate the error on the model so we can find a confidence band on the estimated treatment effect.

The review and analysis of different models in this work demonstrate the importance of using different baseline estimation techniques to evaluate the performance of a DR mechanism. In policy making, specifically, it is vital for the treatment effects to be looked at through different methods as well as using different data sets and trials run on different populations to eliminate biases within the population of study or data set present.

Lastly, in this work specifically and in general in demand response, the goal is to affect consumption patterns in aggregate. The response to a DR pricing mechanism affects both system operators and the grid in aggregate form. The challenge with designing pricing mechanisms is aligning the incentives of all stakeholders: the

consumers and the system operators while aiming for maximum sustainability and stability of the grid.

6.2 Policy Implications

The results from this analysis prove true that a dToU pricing mechanism is effective in lowering consumption during times of high demand. It, additionally, lowers consumers' bills on average while being more cost effective for system operators as the cost of electricity is closer to the marginal cost of generation. Most importantly, tariffs such as dToU help set habits that push towards sustainable goals. Sustainable mechanisms that align incentives and set habits are ever more important as EVs take up a larger fraction of the market. Policy makers can push towards these pricing mechanism becoming more available within the retail landscape to customers in the US. As mentioned, Italy and Ontario, Canada already have ToU pricing present for their residential consumers. Data from these regions can help make a better case for non-static pricing mechanisms.

One important question is whether the socio-economic response as discovered by this work should be utilized by policy makers. When looking at the effects of dToU from a policy standpoint, it's important to look beyond the cost of electricity. What are mental tolls associated with the cost of electricity being significantly higher in certain hours some households simply being unable to shift their consumption outside of those hours? Households with jobs with odd hours — nurses, caregivers, (truck, taxi, rideshare) drivers, security guards, restaurant and bar staff, etc — are negatively impacted by a dToU policy. They may be price-sensitive and may want to take advantage of the incentives but are simply unable to given factors outside of their control.

Another important question is, when looking at treatment effects in aggregate, what disadvantaged subgroups are being overlooked? How might a dToU pricing scheme be regressive or reinforce cycles of harm? For example, it's possible that more rich households are more sophisticated consumers and have the smart appliances to take advantage of a ToU pricing scheme. Additionally, they may have newer and more efficient cooling or heating systems and better insulated walls and windows. Given the above, rich households are more likely to opt-into being priced on a dToU tariff, might change their behavior more, and therefore benefit more.

It is also important to think about ways such pricing schemes need to be regulated in the case of failure modes; for example a price cap may need to be introduced in pricing schemes where the rates are not pre-set or caps on the number of consecutive hours electricity can be priced at a high tariff even if the rates are pre-set. What stakeholder would have to step in in the case a household is unable to pay their utility bill and what happens to that household moving forward? It is important to protect consumers against very high utility bills or being incentivized to shed demand to the point of harm.

Lastly, how does the carbon footprint of the electric sector change as a result of introducing a non-static electricity rate? In a cost-benefit analysis, it is important to consider the long-term environmental impacts of alternative residential electricity rates. For example, the incentives could lead to positive habit setting through positive reinforcement.

6.3 Limitations

This data set included selection bias in both the overall sample and most noticeably in sample under treatment. In this thesis, I used different synthetic control models to

find a mapping between the two groups and remove the bias in the treatment sample but the overall sample was already skewed. For this reason, it's good to use a data set that's blanket applied to everyone (such as Ontario or Italy for ToU pricing) or other trials that assess dToU pricing.

While substantial incentives for recruitment to dTOU helped recruit participants and also made a broader sample possible, minimizing incentives would be preferable due to the buffering effect of incentive payments on price signals and therefore behavior. Recruitment in future trials should consider the possibility of dispensing with, or at least minimizing, incentives and guarantees. Dropping out of the trial would need to be possible, and uninfluenced by payments, in order to study churn rates, which was not possible on the LCL trial. As dTOU becomes less of a novelty, the need for incentives and guarantees should diminish [27, P. 86].

6.4 Future Work

In addition to what was outlined in sections **??** and **??**, there are some more angles this work can take, both using the data set at hand and otherwise. Using the same data set, it would be interesting to use a more complex imputation function to see if it would be possible to complete the data matrix with less error introduced while keeping the temporal trends present. As a reminder the current data is being sliced as the majority of the data was missing in 2012 as shown in figure 4-6.

The data set has two Acorn groups: one that is more granular and includes 17 classifications and one that is less granular and has 3 classifications and is the one I used. It would be interesting to redo some of the analyses using the less granular Acorn classification to see if there is a more gradual response to the treatment.

It would be interesting and would add to the confidence on these models to work

with a data set that has fewer limitations, includes a longer period of time, and has more features. The added features could help truly see the power of some of the machine learning models such as TFT, Prophet, and random forest regression.

6.4.1 Statistical Implications of Modeling Using Noisy Data

There is a debate in statistics that we model behavior using uncertain data only to find what the uncertainty in the data is — how effective the treatment has been. This is observed in this work as well. Sensitivity analysis — which quantifies how much model output values are affected by changes in model input values — can help ground the analysis. Similarly, an uncertainty analysis aims to shed light on the effect of the noise in the data on model outputs by taking a set of randomly chosen input values (which can include parameter values), passing them through a model (or transfer function) to obtain the distributions (or statistical measures of the distributions) of the resulting outputs. Such metrics can better inform of the sensitivity of our models to the input data.

Bibliography

[1] Low carbon london program website.

[2] Regional transmission organization (north america).

[3] Acorn user-guide, 2014.

[4] Acorn technical guide, 2021.

[5] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505, 2010.

[6] Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: A case study of the basque country. *American economic review*, 93(1):113–132, 2003.

[7] Anish Agarwal, Muhammad Jehangir Amjad, Devavrat Shah, and Dennis Shen. Model agnostic time series analysis via matrix estimation, 2018.

[8] Neil Lessem Dean Mountain Frank Denton Byron Spencer Chris King Ahmad Faruqui, Sanem Sergici. Analysis of ontario's full scale roll-out of tou rates – final study, 2016.

[9] Perez-Arriaga Schneider von Scheidt Burger, Knittel. The efficiency and distributional effects of alternative residential electricity rate designs. *The Energy Journal*, 41(1):199–240, 2020.

[10] Jesse Burkhardt, Kenneth Gillingham, and Praveen K Kopalle. Experimental evidence on the effect of information and pricing on residential electricity consumption. (25576), February 2019.

[11] California Energy Commission California Independent Systems Operator, California Public Utilities Commission. Preliminary root cause analysis: Mid-august 2020 heat storm.

[12] M. A. Crew and P. R. Kleindorfer. Reliability and public utility pricing. *The American Economic Review*, 68(1):31–40, 1978.

[13] 2014 by POWER Dec 18. Ferc order 745 and the epic battle between electricity supply and demand, Dec 2014.

[14] Inc. EnerNOC. The demand response baseline. 2011.

[15] Ito. Do consumers respond to marginal or average price? evidence from nonlinear electricity pricing. *American Economic Review*, 104(2):537–563, 2014.

[16] Matthew S. Johnson. Towards machine learning-based demand response forecasting using smart grid data, 2021.

[17] Paul Joskow and Jean Tirole. Retail electricity competition. *The RAND Journal of Economics*, 37(4):799–815, 2006.

[18] Richard J. Pierce Jr. A primer on demand response and a critique of ferc order 745. 2011.

[19] D.S. Kirschen. Demand-side view of electricity markets. *IEEE Transactions on Power Systems*, 18(2):520–527, 2003.

[20] D.S. Kirschen and G. Strbac. *Fundamentals of Power System Economics*. Wiley, 2004.

[21] Teuvo Kohonen. volume 30. Springer.

[22] Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting, 2020.

[23] Alex Whitney Jelena Dragovic James Schofield Matt Woolf Goran Strbac Mark Bilton, Richard Carmichael. Accessibility and validity of smart meter data.

[24] NASA. Power access data viewer.

[25] Saehong Park, Seunghyoung Ryu, Yohwan Choi, Jihyo Kim, and Hongseok Kim. Data-driven baseline estimation of residential buildings for demand response. *Energies*, 8(9):10239–10259, 2015.

[26] UK Parliament. Climate change act 2008.

[27] Matt Woolf Mark Bilton Ritsuko Ozaki Goran Strbac Richard Carmichael, James Schofield. Residential consumer attitudes to time-varying pricing.

[28] Nerea Ruiz, Iñigo Cobelo, and Jose Oyarzabal. A direct load control model for virtual power plant management. *IEEE Transactions on Power Systems*, 24:959–966, 2009.

[29] David Salinas, Valentin Flunkert, and Jan Gasthaus. Deepar: Probabilistic forecasting with autoregressive recurrent networks, 2019.

[30] Ian Schneider, Mardavij Roozbehani, and Munther Dahleh. An online learning framework for targeting demand response customers. *IEEE Transactions on Smart Grid*, 13(1):293–301, 2022.

[31] James Schofield. *Dynamic time-of-use electricity pricing for residential demand response: Design and analysis of the Low Carbon London smart-metering trial.* PhD thesis, 2015.

[32] Rajesh Subbiah, Anamitra Pal, Eric Nordberg, Achla Marathe, and Madhav Marathe. Energy demand model for residential sector: A first principles approach. *IEEE Transactions on Sustainable Energy*, PP:1–1, 02 2017.

[33] Sean J. Taylor and Benjamin Letham. Forecasting at scale. *PeerJ Prepr.*, 5:e3190, 2017.

[34] The Pew Charitable Trusts. America's electric grid: Growing cleaner, cheaper and stronger. https://www.pewtrusts.org/en/research-and-analysis/reports/2015/10/americas-electric-grid-growing-cleaner-cheaper-and-stronger, October 27, 2015 (accessed March 29, 2021).

[35] Hsiang-Fu Yu, Nikhil Rao, and Inderjit S Dhillon. Temporal regularized matrix factorization for high-dimensional time series prediction. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.