Peaking Interest: How awareness drives the effectiveness of time-of-use electricity pricing

Brian C. Prest*

July 3, 2017

Abstract

I apply and extend new machine learning methods to identify heterogeneity in the effects of time-of-use (TOU) electricity pricing and information provision on residential electricity consumption behavior in an experiment from the Republic of Ireland. Most importantly, the effect of time-of-use pricing on peak energy consumption is 4.5 times larger for households who are aware of the change in their pricing structure versus those who are not (-10% versus -2.3%). Households with low baseline energy usage do not reduce their peak consumption on average. Information provision amplifies the effects on peak consumption, with an in-home electricity monitor nearly doubling the effectiveness of TOU pricing (-15% with one versus -8% without). Cross-validation finds that no other observables—or permutations thereof—are robustly related to treatment effect heterogeneity (conditional on the above factors of awareness, baseline consumption, and information treatment). This includes considering potential heterogeneity on more than 150 observables encompassing socio-demographic characteristics, attitudes towards energy and environmental issues, housing attributes, and household appliance characteristics. In addition, larger price increase do not appear to induce significantly larger responses. Awareness is not reliably predictable even with rich information about observable household characteristics, suggesting that targeting policy on awareness is unlikely to prove fruitful, although targeting on baseline consumption could be effective. These results suggest that the significant attention economists pay towards fine-tuning retail prices may be less important than getting consumers to pay attention in the first place.

1 Introduction and Motivation

Electricity consumers typically pay a constant price for electricity that is not tied to the marginal cost of generation, which varies significantly with demand and time of the day. Recent advances in smart metering technology now allow for charging consumers time-varying prices, raising the possibility of flexible demand that can respond to price signals in real time. The benefits of dynamic pricing are threefold. First, the mismatch between constant retail prices and time-varying marginal costs is economically inefficient, encouraging consumers to over-consume during high-cost peak periods and under-consume during low-cost off-peak periods. Second, time-of-use (TOU) pricing can reduce peak load for the grid as a whole, mitigating the need to invest in costly extra "peaking" generation capacity to serve load when demand is high. Third, recent studies (e.g.,

[*] Duke University, brian.prest@duke.edu, https://sites.duke.edu/brianprest/. I appreciate input from seminar participants at Duke University and the Summer 2017 AERE conference. Special thanks to Faraz Usmani for paper title suggestions. I am also grateful to to the Irish Social Science Data Archive (ISSDA) for providing access to the data used in this study (accessed via the Irish Social Science Data Archive - www.ucd.ie/issda). Any errors are my own. ISSDA bears no responsibility for this study's analysis or interpretation thereof. Copyright and all other intellectual property rights in the data and associated documentation are vested in the original data creators. The online appendix can be found at https://www.dropbox.com/s/pki7uicewsjmjma/Prest_TOU_Pricing%20-%20Online%20Appendix.pdf?dl=0

Finn and Fitzpatrick 2014; Fripp 2016) demonstrate that time-varying pricing can help accommodate the integration of renewable generation by better aligning the timing of demand with that of intermittent supply, thereby reducing the costs of meeting environmental goals in the power sector.

For all of these reasons, understanding how and under what conditions consumers are likely to respond to time-of-use pricing is of growing importance for electricity and environmental policy. This paper applies and extends new machine learning methods to estimate heterogeneous responses to TOU pricing to consider the conditions under which such pricing is more or less effective.

A growing body of evidence suggests that time-varying pricing can significantly reduce energy usage during peak periods (see, e.g., Ito, Ida, and Tanaka 2015, Wolak 2011). The literature has also suggested that information provision can help amplify these effects (e.g., Jessoe and Rapson 2014). Nonetheless, the magnitudes of these responses have often been smaller than hoped. One explanation for this is that existing studies typically do not systematically assess the heterogeneity in these effects. For example, a modest average response may simply represent large responses for some consumers, offset by negligible effects for others.

Understanding this heterogeneity can shed light on why policies may underperform on average, thereby suggesting lessons for improving policies in the future. For example, by identifying particularly important subgroups that are more or less responsive, policies can be targeted towards households for whom the policies are more likely to be effective. Heterogeneous effects can also inform policy design, for example by suggesting ways to make pricing policies more salient to consumers (and whether such an effort would be worthwhile).

Researchers rarely estimate heterogeneous effects of TOU pricing in a systematic fashion. Typically, only one or two potential dimensions of heterogeneity are considered, rather than the many possible dimensions and their interactions. For example, a researcher may first estimate an average treatment effect, and then consider how the effect varies by educational status (as in Di Cosmo, O'Horaa, and Devitt 2015). Big data makes many such comparisons possible, which allows researchers to explore many dimensions of treatment effect heterogeneity. However, rich datasets also raise the two problems: the curse of dimensionality and multiple hypothesis testing. To overcome these problems, this paper applies a modern machine learning technique that provides a parsimonious way to estimate sources of heterogeneity in treatment effects that are robust to out-of-sample validation.

I consider heterogeneity in treatment effects of an experiment in which Irish households were randomly allocated to either a control group or one of sixteen treatment groups receiving a combination of energy usage information and time-of-use pricing. The time-of-use pricing component involved higher prices during peak periods (5pm-7pm), significantly lower prices at night (11pm-8am), and slightly lower prices at other times

(8am-5pm and 7pm-11pm, referred to as "daytime" hours).[1] The information provision component involved providing households with some combination of an energy usage statement, more frequent billing, and an in-home display (IHD) showing real-time electricity prices and usage.

I apply and extend the tree-based machine learning method from Athey and Imbens (2016) to estimate the key sources of heterogeneity in treatment effects of TOU pricing. I extend their algorithm to a difference-in-differences setting and with multiple treatment groups, contributing to a new and developing area in machine learning for causal inference (see, for example, Athey, Tibshirani, and Wager 2016; Deng, Zhang, Chen, Kim, and Lu 2016). While my extension of this algorithm is new, its implementation is straightforward and can be executed using standard out-of-the-box packages for regression trees, such as `tree` or `rpart` in R, or alternatively in the `causalTree` package, requiring only moderate data pre-processing.

Most importantly, the algorithm reveals that the treatment effect of peak pricing is 4.5 times larger for households who are aware of the change in their pricing structure versus those that are not (-10% versus -2.3%, latter of which is statistically indistinguishable from zero). This suggests that the salience of price changes is key to its effectiveness, consistent with Chetty, Looney, and Kroft (2009). It also suggests that substantial additional savings could be achieved by ensuring that households understand their tariff structure.

Beyond awareness, the next-largest source of heterogeneity is baseline electricity consumption, whereby households with very low baseline energy use did not reduce peak consumption (and may actually increase it in some cases), suggesting that time-of-use pricing should target larger consumers. The next largest sources of heterogeneity are the different information provision treatments, with larger information stimuli leading to larger treatment effects. The difference is substantial, with the most effective information stimulus (the in-home display (IHD)) delivering an average treatment effect of -15%, nearly twice as large as the weakest information stimulus -8% (a bi-monthly energy usage statement). This suggests that information provision is crucial when implementing time-of-use pricing, even among those who report being aware of the policy.

Conditional on these sources of heterogeneity (awareness, baseline consumption, and information treatment), no other sources are robustly related to treatment effect heterogeneity, despite having rich data on households' family characteristics, house characteristics, appliances and electronics, and more. This suggests that, conditional on a household's baseline consumption, little additional information is necessary to predict treatment effects.

In addition, the treatment effects are not significantly larger for treatment groups experiencing larger price increases. In particular, the point estimates of the effects scale less-than-proportionally with the size of the

[1] Time-of-use pricing was only active on non-holiday weekdays. For this reason, I only include these days in my analysis, unless otherwise stated. In the online appendix, I also conduct a placebo test for a treatment effect during weekends and holidays, finding little to no effect.

price change, implying a non-constant demand elasticity. This implies diminishing returns to larger price increases, consistent with Ito, Ida, and Tanaka (2015).

While there is significant heterogeneity in the treatment effects during peak periods, there is no detectable heterogeneity during non-peak periods. During nighttime hours, the treatment increased usage about 1.8%, whereas during daytime hours it reduced usage by about 2.2%.[2] The latter result may appear somewhat anomalous, as households reduced daytime (non-peak) consumption despite lower prices. This suggests that attentiveness and adjustment costs may lead households to reduce daytime (off-peak) consumption in anticipation of peak prices.

In contrast to standard econometric approaches, the machine learning algorithm produces a parsimonious set of factors determining heterogeneity in treatment effects during peak periods: awareness, baseline consumption, and information provision. An naive alternative would be to test all possible covariate interactions for statistical significance. In this data, such an approach produces a larger set of covariates apparently associated with treatment effect heterogeneity, with approximately 7% of all such coefficients“statistically significant” at the 5% level. The Athey-Imbens algorithm determines which coefficients represent true sources of heterogeneity that are robust to out-of-sample validation, as opposed to spurious relationships. The value of the machine learning method is that it culls the spurious estimates of heterogeneity while retaining those that withstand out-of-sample validation.

Given that awareness is the key driver of heterogeneity, I also explore what characteristics of households best predict whether they will understand — and hence respond to — the treatment. I find that observable household characteristics have little predictive power for consumer awareness, despite having over one hundred such covariates spanning demographics, attitudes toward electricity use, and physical house characteristics. This suggests that better explanation of time-of-use pricing mechanisms—not targeting awareness—is likely to be more successful at increasing the effectiveness of time-of-use pricing. It also suggests that the significant attention economists pay towards fine-tuning retail prices is less important than getting consumers to pay attention in the first place.[3]

This study makes two primary contributions. First, it identifies the key sources of treatment effect heterogeneity in time-of-use pricing using the latest techniques from machine learning, based on Athey and Imbens (2016). The algorithm retains only the dimensions of heterogeneity that are robust to out-of-sample validation, and it eliminates those that are not. This reveals that awareness of the policy is key to program effectiveness, and few other dimensions beyond that matter for observable heterogeneity, including the degree of the price increase.

[2] The t-statistics are 1.6 and -2.9, respectively. Overall, the program reduced mean consumption by 2.3% ($p = 0.02$)

[3] A caveat to this is that this result is based on an experiment of Irish residential consumers without widespread automation technology. It is possible that other types of consumers (commercial or industrial consumers, or consumers in other countries) may be more responsive. It is also possible that households with more automation technology, such as thermostats that can be programmed to respond to real-time prices, would be more sensitive to the price levels.

Second, it extends the Athey and Imbens (2016) algorithm to settings with multiple treatment groups. This extension allows for using that algorithm to find heterogeneous treatment effects not only across household observables but also across treatment groups. This is important in settings such as this one, where some treatment dimensions (i.e., information) significantly mattered for treatment effect heterogeneity while other dimensions (i.e., pricing magnitude) did not.

2 Literature

The economic literature has generally found that dynamic electricity pricing can reduce consumption during peak hours. Faruqui and Palmer (2012) provide a survey of the evidence as of 2012, finding that pricing incentives are generally effective, but larger price increases lead to diminishing returns. They also find that information provision generally enhances the effectiveness of the policies. Wolak (2011) finds that the consumers respond to comparable "critical peak pricing" (CPP) and hourly TOU pricing policies in a similar manner.[4] In more recent evidence, Ito, Ida, and Tanaka (2015) find that CPP has more persistent effects on peak consumption than behavioral nudges do. Herter and Wayland (2010) also finds significant reductions during CPP events, particularly for larger consumers and those in cool climate zones, but finds that consumers did not respond more to larger price increases.

Blonz (2016) finds that commercial and industrial consumers in California respond to CPP pricing, but only in hot inland regions and not in milder coastal regions. Thorsnes, Williams, and Lawson (2012) also finds significant effects from an experiment in New Zealand, but only during the winter months when demand is highest. That study also explored some heterogeneity in the effect, finding it varied with household size and whether its water is heated with electricity. Torriti (2012) finds significant load shifting in response to TOU pricing, but finds that this actually increased total consumption due to higher usage during low-price periods. Cappers, Spurlock, Todd, Baylis, Fowlie, and Wolfram (2017) discusses the importance of default options in driving electricity consumption in response to TOU pricing.

This literature often finds that information provision helps increase the effectiveness of dynamic pricing. For example, Jessoe and Rapson (2014) find that consumers given IHDs exhibited stronger and more consistent effects on consumption during CPP events than did consumers without IHDs. Ivanov, Getachew, Fenrick, and Vittetoe (2013) similarly find that IHDs amplified the effectiveness of CPP events. Bollinger and Hartmann (2016) also find that IHDs amplified the impacts of dynamic pricing, but they further find that automation of electricity consumption through programmable thermostats has even larger effects. Together, the literature

[4] There are multiple kinds of dynamic electricity pricing. One is "critical peak pricing" (CPP), in which consumers face very large price increases (e.g., 300% increases) but only during rare and extreme events (e.g., perhaps few times per year, on very hot days). Another is time-of-use (TOU) pricing under which consumers' prices routinely vary from hour to hour. To avoid overly complicated pricing schedules, usually only two or three rates are applied (e.g., a low rate at night, a moderate rate during most daytime hours, and a high rate during peak hours).

suggests that information provision can help increase the effectiveness of dynamic pricing, and perhaps even more can be done, such as automation or targeting.

While these studies generally find that time-varying pricing can reduce peak consumption, they typically do not systematically assess the significant heterogeneity in these effects. A systematic assessment of heterogeneity can help identify the underlying causal mechanisms and the conditions under which pricing policies are most effective, which in turn can improve policy targeting and design. This paper contributes to the literature on time-varying electricity pricing by assessing many possible dimensions of treatment effect heterogeneity, while avoiding multiple hypothesis testing concerns through a modern machine learning algorithm.

A recent working paper (Burlig, Knittel, Rapson, Reguant, and Wolfram 2017) is relevant to this study because it also applies machine learning to estimating causal effects from high-frequency electricity consumption data. That study uses Lasso-based projections of counterfactual electricity consumption to estimate the impact of energy efficiency investments, finding impacts less than half as large as projected by engineering estimates. While that paper studies a different question, it illustrates the value of applying machine learning techniques to estimate causal effects in electricity interventions.

Finally, a small number of studies have explored the data from the same experiment in Ireland, most of which are yet to be published (as of summer 2017). Carroll, Lyons, and Denny (2014) assess the extent to which the treatment increased participants' self-reported knowledge about their energy consumption and how to reduce it; they find that while the program did increase participants' knowledge, such increases were not correlated with actual reductions in consumption. McCoy and Lyons (2016) find that the treatment actually made participants *less* likely to install energy-saving technologies, suggesting unintended side effects. Pon (2015) finds that the information provision reduced peak consumption but that this effect decayed over time. Finally, Di Cosmo, O'Horaa, and Devitt (2015) find significant treatment effects and find that more educated households did not respond more overall, but perhaps made better use of the information treatments.

3 Data

3.1 Description of the Program

The data for this study comes from the smart meter Consumer Behavior Trial (hereafter, the "trial") run during 2009-2011 by the Commission for Energy Regulation (CER) in the Republic of Ireland. The experiment was designed to assess the effectiveness of various time-of-use pricing structures and demand-side management stimuli on consumer electricity consumption. In the experiment, CER recruited a nationally representative sample of Irish households. Participation was voluntary, with a financial incentive of €25 per survey completed.

In addition, treated households received two credits to offset the higher expected costs from being on a variable tariff.[5] All credits were paid outside of the treatment period to avoid income effects.[6]

Households were randomly assigned to either treatment and control groups.[7] Smart meters were installed in both treatment and control houses. Before the treatment began, baseline electricity consumption data was collected at the half-hourly resolution from July through December of 2009.[8] The treatment period was January to December 2010. Households who were assigned to be treated were allocated to one of four time-of-use tariffs and also one of four demand-side management stimuli. This four-by-four treatment structure implies sixteen distinct treatment groups.[9]

The time-of-use pricing schedules ranged from somewhat peaked (12 cents per kWh at night vs. 20 cents during peak hours (all values in €)) to very strongly peaked (9 cents vs. 38 cents, respectively). These pricing schedules are illustrated in Figure 1.[10] The pricing structures were designed with the goal that the "average" participant who did not change their consumption behavior would not face higher bills on average.

For the information stimuli, all treated households received a refrigerator magnet and sticker explaining the time-of-use pricing scheme and an energy usage statement along with their bill. The first information stimulus simply involved providing the energy usage statement along with the household's bi-monthly bill (that is, every other month). The second information stimulus involved billing consumers more frequently (every month). The third information stimulus billed bi-monthly but provided households with an in-home electricity monitor displaying both real-time and historical information about energy usage and prices.[11] The final stimulus involved providing the bi-monthly billing statement plus an "Overall Load Reduction" (OLR) incentive. The OLR stimulus offered households a €20 bonus if they could reduce their baseline consumption by 10% or more. Control households experienced no change in their electricity rate (a constant 14.1 cents) or billing frequency (bi-monthly).

The program also included pre- and post-trial surveys, for which participants received €25 per survey to complete. These involved hundreds of questions covering many topics, including socio-demographic

[5] These credits were €30, €50, €70, or €90 to households in the A, B, C, or D tariff groups, respectively.

[6] For a detailed summary of the trial, see Commission for Energy Regulation (2011b).

[7] The random allocation was designed to ensure balance on covariates across experimental cells. Namely, households were classified into one of several "profiles" based on a principal component analysis of energy usage and survey data. Then, households in each profile were randomly assigned to treatment/control groups. In addition, some participants were moved to different cells after the initial allocation to improve balance between cells.

[8] Households were told of their treatment/control assignment late during this benchmark period (December 2009) to avoid potential effects on baseline behavior.

[9] There was also 17th treatment group with a relatively small number of participants who faced a special weekend tariff and no information stimulus. I drop these households from my analysis because they faced a fundamentally different treatment.

[10] Peak tariffs were 20, 26, 32, and 38 cents per kWh for tariffs A, B, C, and D, respectively. The corresponding rates during off-peak daytime periods were 14, 13.5, 13, and 12.5 cents. At night, they were 12, 11, 10, and 9 cents. The control group faced a flat price of 14.1 cents for all hours.

[11] See online appendix for pictures illustrating the monitor and energy usage statements.

characteristics, attitudinal questions, physical attributes of the home, and self-reported expectations about energy use.[12]

Unfortunately, some households did not complete the surveys. For my analysis, I only include households that did so (74% of the full sample). However, for households that did not, I can still observe their energy consumption and their treatment assignment. This allows me to test whether the average treatment effect is significantly different for those who did not answer the survey, and I find that it is not.[13] This suggests that the survey response rate is not strongly biasing my treatment effect results.[14]

The treatment/control distribution of the 3,006 remaining households is shown in Table 1.[15]

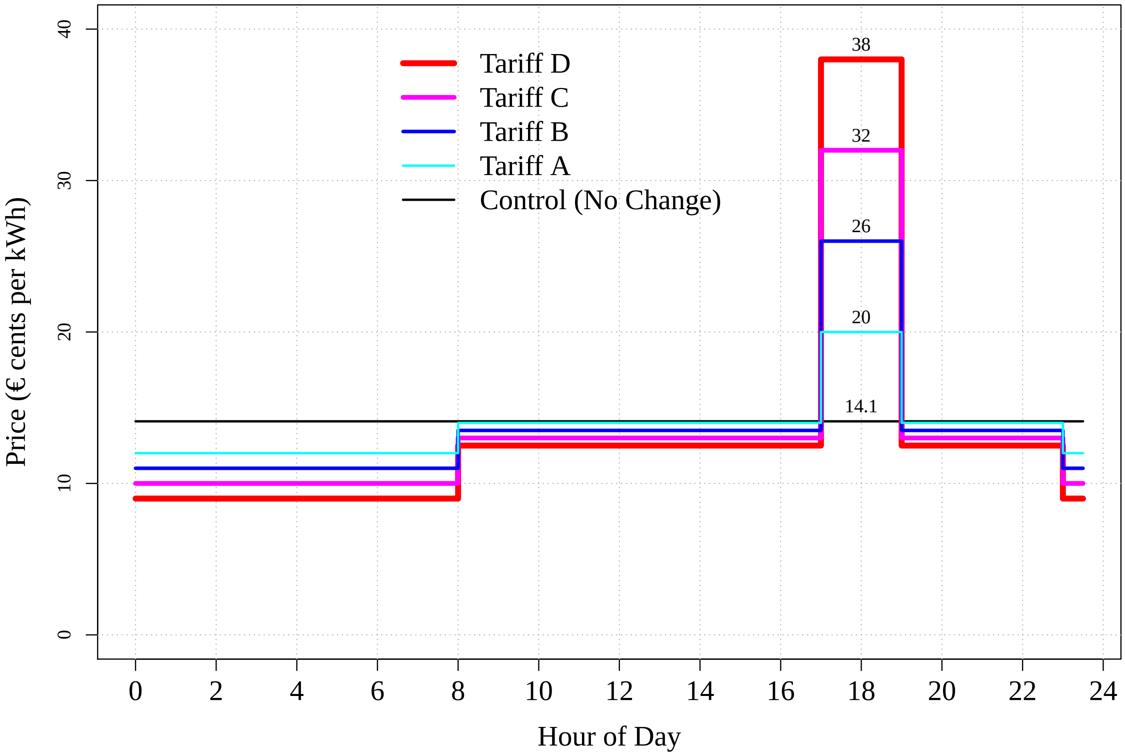


Figure 1: Time-of-Use Pricing Structures

[12] The pre- and post-trial surveys include 377 questions combined. Full lists of these questions are available from ISSDA, here and here. Not every household was asked every question however. For example, questions about the in-home electricity monitors were not asked of households who did not receive a monitor. Survey questions not answered by all households must be dropped from the analysis, leading to the approximately 150 variables I use. See the online appendix for a list of the survey questions included in the analysis.

[13] The average effect on peak consumption is 8.3% for the full sample, compared to 8.9% for those who completed both surveys (not significantly different; $p = 0.38$).

[14] In addition, about 21% of households were lost to attrition, the vast majority of which (83%) was due to a change of supplier. The attrition rate was only slightly higher for control households (25%) than it was for treatment households (20%). An investigation into the reasons for changes in supplier by CER found that "[a]mong attritors who changed supplier [who represent the vast majority of attritors], the reason for the switch appears to have been independent of any potential impact of the Trial. None of these switchers stated that the tariffs and or technology were a factor in this decision. 5% stated that the consumption reports and other information contributed to the decision." (Commission for Energy Regulation 2011a). I also drop a small number of households because they have extended periods of zero consumption, as these are likely to be frequently uninhabited vacation homes.

[15] Fewer households were assigned to tariffs B and D by design based on CER's power analysis.

Table 1: Treatment and Control Group Assignments

	Bi-monthly Bill & Energy Usage Statement	Monthly Bill & Energy Usage Statement	Bi-monthly Bill, Energy Usage Statement, and In-Home Display (IHD)	Bi-monthly Bill, Energy Usage Statement, and Overall Load Reduction (OLR) Incentive	Control	**Total**
Tariff A	195	216	205	216	0	**832**
Tariff B	80	87	72	81	0	**320**
Tariff C	222	217	202	213	0	**854**
Tariff D	80	87	78	77	0	**322**
Control	0	0	0	0	678	**678**
Total	**577**	**607**	**557**	**587**	**678**	**3,006**

3.2 Treatment/Control Balance

While treatment/control groups were assigned randomly, the groups did not end up being perfectly balanced. Table 2 shows the group means and test statistics for balance for a variety of observable pre-trial characteristics between the control and treatment groups. The top panel of that table includes all variables that were found to be significantly different between the treatment and control groups. The bottom panel shows a selection of the variables with non-significant differences.

Some of the variables that are statistically different between treatment and control groups are worrying because they are directly related to energy consumption. Baseline night-time consumption (variable #9) is larger for treatment households (0.15 kWh per half hour) than for control households (0.14 kWh). Peak and daytime consumption are also somewhat higher, although not statistically different ($p = 0.12$ and $p = 0.07$, respectively).

In total, 122 pre-trial variables are tested, of which 14 (11%) are significantly different for treatment and control at the standard 5% level, although none are significant when using the Bonferroni correction for multiple hypothesis testing. If the groups were perfectly balanced and all variables were independent, we would expect approximately 5% of the variables tested to be significant at this level. One may, then, be concerned that the larger share, 11%, of the variables is significant. However, many of these variables are strongly correlated with each other. For example, variable #3, “Number of Electronics” is equal to the sum of all electronics in the household, including variables #2 (number of large televisions) and #11 (number of desktop computers). Similarly, variable #6, “Number of Residents”, is perfectly collinear with the number of adults (#16) in the home, the number of children (#12), and the indicator variable indicating whether any children live in the home (#5). These variables are perfectly collinear, so interpreting them as independent sources of imbalance would be inappropriate.

To adjust for this perfect collinearity, I also estimate a linear probability model[16] of treatment status on pre-treatment household observables. The results of this estimation are shown in Table 3. Here, the observables do little to predict treatment status, as only 2% of the variables tested are statistically significant at the 5% level, and again none are significant with the Bonferroni correction. In particular, the coefficients on the baseline consumption variables are small and insignificant, suggesting that energy consumption is balanced *conditional on household observables*.[17] This suggests the treatment and control groups are balanced on average over the full set of observables, if not on each specific one taken individually.

While the small imbalances between the groups are not generally statistically significant, it is worth considering differences on a key variable of interest: average peak electricity consumption during the baseline period (variable #15 in Table 2). Baseline peak consumption is slightly larger for the treatment group (0.44 kWh per half hour) than for the control group (0.42 kWh). This amounts to an approximately 5% difference. While the difference is not statistically significant, its magnitude is non-trivial relative to typical treatment effect sizes. For this reason, I pursue a difference-in-differences strategy to net out time-invariant factors that are different between the treatment and control groups. Given the differences in treatment in control baseline consumption, a single treatment/control difference would be badly biased.[18] This highlights the importance of the difference-in-differences approach.[19]

4 Average Treatment Effects Using Difference-in-Difference

In this section, I use a difference-in-differences approach to estimate average treatment effects and highlight the sources of identification. In the subsequent section, I explore the heterogeneity in these effects.

4.1 Half-Hourly Average Treatment Effects

I estimate the average treatment effect for each 30-minute interval of the day in a difference-in-differences framework. I do not distinguish between the sixteen different treatment groups for the moment, as I will estimate that dimension of heterogeneity in subsequent sections. For each of the 48 half-hour periods of the

[16] A logistic regression produces the exact same set of statistically significant coefficients as the linear probability model and hence has the same interpretation. I present the linear probability model because it is easier for the reader to interpret.

[17] Further, performing model selection by Lasso concludes that all or nearly all covariates are unrelated to treatment, depending on the criteria used to choose the tuning parameter. As discussed at length in section 5.4 below, there are two metrics are commonly used for choosing tuning parameter in the Lasso: the RMSE-minimizing lambda and the 1 standard error lambda. Here, the conservative 1 standard error rule finds that no covariates are related to treatment status, and the RMSE-minimizing lambda finds that only the number of televisions and social class variables are related.

[18] For example, if the treatment effect were actually -7%, then estimating the treatment effect by subtracting treatment and control means would recover a severe underestimate of -2% ($= -7\%$ treatment effect $+ 5\%$ bias).

[19] In addition, perfect balance is not required for the Athey and Imbens (2016) estimator I employ in the subsequent section. That estimator instead requires only the weaker assumption of unconfoundedness (i.e., that *conditional on observables* treatment is uncorrelated with the outcome variable). Further, I show in the online appendix that my results are robust to using propensity score weighting to correct for imbalance on observables.

Table 2: Treatment/Control Balance: Covariate t-tests

	Variable	Control Mean	Treatment Mean	t-statistic	p-value
	Unbalanced Variables ($\alpha < 0.05$)				
1	Employment status: Retired (Indicator)	0.38	0.31	3.40	0.001
2	Number of Large Televisions (21+ inch)	1.19	1.31	-3.36	0.001
3	Number of Electronics	3.74	4.04	-3.03	0.003
4	Age Group: 65+ (Indicator)	0.28	0.23	2.79	0.01
5	Has Children Under 15 in Home (Indicator)	0.23	0.28	-2.78	0.01
6	Number of Residents	2.60	2.76	-2.65	0.01
7	Social Class: AB (Highest) (Indicator)	0.12	0.15	-2.58	0.01
8	Education: Primary only (Indicator)	0.15	0.11	2.51	0.01
9	Baseline Average Consumption (Night Hours)	0.14	0.15	-2.42	0.02
10	Internet Access in Home (Indicator)	0.66	0.71	-2.24	0.02
11	Number of Desktop Computers	0.48	0.53	-2.18	0.03
12	Number of Children Under 15 in Home	0.43	0.52	-2.11	0.04
13	Housing Status: Own with Mortgage (Indicator)	0.35	0.40	-2.09	0.04
14	Others in Household Use Internet Regularly (Indicator)	0.53	0.57	-2.01	0.04
	Selected Balanced Variables ($\alpha \geq 0.05$)				
15	Baseline Average Consumption (Peak Hours)	0.42	0.44	-1.85	0.07
16	Number of Adults in Home	2.16	2.24	-1.74	0.08
17	Cook stove type: Electric (Indicator)	0.72	0.69	1.62	0.10
18	Number of Laptop Computers	0.65	0.71	-1.61	0.11
19	Baseline Average Consumption (Day Hours)	0.29	0.30	-1.56	0.12
20	Unemployed, not seeking job (Indicator)	0.03	0.04	-1.52	0.13
21	Home Heat: Solid Fuel (Indicator)	0.29	0.26	1.46	0.14
22	Interested in changing energy use for environment*	1.38	1.34	1.41	0.16
23	Female (Indicator)	0.47	0.50	-1.02	0.31
24	Education: Secondary to Certificate (Indicator)	0.16	0.17	-0.86	0.39
25	Satisfied with billing frequency*	2.84	2.86	-0.47	0.64
26	Expect to Choose More Efficient Appliances*	1.34	1.35	-0.35	0.73
27	Number of Immersion Water Heaters	0.77	0.77	-0.15	0.88
28	Home Style: Terraced (Indicator)	0.14	0.14	-0.14	0.89
29	Number of Washing Machines	0.99	0.99	-0.04	0.97
30	Unemployed, seeking job (Indicator)	0.04	0.04	-0.01	0.99
	Observations	3,006			
	Number of Variables Tested	122			
	Number of Variables Not Shown	92			
	Number of Variables Significant (5% level)	14			
	Share of of Variables Significant (5% level)	11.5%			

Notes: Due to space limitations, only a subset of the 122 tested variables are shown. The suppressed variables are all statistically insignificant at the 5% level. The full set of t-tests is available upon request. The variables are presented in ascending order of p-value. Baseline Average Consumption variables units are kWh per 30 minute interval. This table was generated using the `stargazer` package (Hlavac 2015) for R.

* These variables featured numeric responses, where respondents reported on a 1-5 scale to what extent they agree (1) or disagree (5) with the statement.

Table 3: Treatment/Control Balance: Linear Probability Model of Treatment on Covariates

	Dependent variable:
	Treated (Indicator)
Baseline Average Consumption (Peak Hours)	0.03
	(0.07)
Baseline Average Consumption (Night Hours)	0.14
	(0.15)
Baseline Average Consumption (Day Hours)	−0.09
	(0.12)
Employment Status: Retired (Indicator)	−0.04
	(0.05)
Number of Large Televisions (21+ inch)	0.02^{**}
	(0.01)
Age Group: 65+ (Indicator)	0.13
	(0.11)
Has Children Under 15 in Home (Indicator)	0.06
	(0.04)
Number of Adults in Home	0.01
	(0.01)
Social Class: AB (Highest) (Indicator)	−0.03
	(0.08)
Education: Primary Only (Indicator)	−0.07
	(0.04)
Internet Access in Home (Indicator)	0.01
	(0.02)
Number of Desktop Computers in Home	0.01
	(0.02)
Number of Children Under 15 in Home	−0.01
	(0.02)
Others in Household Use Internet Regularly (Indicator)	−0.003
	(0.02)
Cook stove type: Electric (Indicator)	-0.13^{**}
	(0.06)
Constant	0.43
	(0.28)
Observations	3,006
R^2	0.03
Adjusted R^2	-0.01
F Statistic	0.76
Number of Covariates	109
Number of Covariates Not Shown	94
Number of Covariates Significant (5% level)	2
Share of Covariates Significant (5% level)	1.8%

*p<0.1; **p<0.05; ***p<0.01

Notes: Standard errors shown in parentheses. Due to space limitations, only a subset of the 109 included covariates are shown in this table. The full results are available upon request. All suppressed covariates have statistically insignificant coefficients (at the 5% significance level). There are fewer covariates in this table than in the t-tests in Table 2 because of perfect collinearity (e.g., number of residents equals the sum of number of adults and children). Baseline Average Consumption variables units are kWh per 30 minute interval. This table was generated using the `stargazer` package (Hlavac 2015) for R.

day, I separately estimate the following specification for each half hour of the day $j \in \{1, 2, ..., 47, 48\}$:

$$\ln(Y_{i,j,t}) = \beta_j W_{i,t} + \alpha_i + \lambda_m + \epsilon_{i,j,t}. \tag{1}$$

$Y_{i,j,t}$ is the electricity consumption by household i in during time window t (a half-hour, e.g., 5:00pm-5:30pm on March 9, 2010). $W_{i,t}$ is an indicator equal to one for treatment households during the treatment period, and is zero otherwise.[20] α_i and λ_m are household and month-of-sample fixed effects.[21] The parameters of interest are β_j, particularly during the peak periods, when prices are highest. These represent the local average treatment effect (LATE) for each hour of the day.[22] The point estimates for β_j, along with 95% confidence intervals, are shown in graphical form in Figure 2.

The results indicate a substantial reduction in energy consumption during peak hours, consistent with the substantially higher peak-period price (as shown in Figure 1). Similarly, low prices at night led to moderate increases in electricity consumption at the beginning of the night time price period (11pm-12:30am), but this effect largely disappears for the remainder of the night hours.

During non-peak daytime hours, there appears to be a moderate reduction in energy usage during some hours, although this effect is only statistically significant for 7 of the 26 half-hour-long periods. This is interesting however because time-of-use prices during these hours were *lower* than the prices both for the control group and for the treatment group before the trial was implemented (see Figure 1). Taken literally, this would imply an upward-sloping demand curve.

One explanation for this phenomenon is that adjustment costs prevent households from changing their consumption on an hourly basis during workdays. Under this theory, households know that peak prices will be charged late in the day, but they will not necessarily be at home at 5:00pm to turn off devices. Anticipating this, they adjust appliances before leaving home in the morning. This is consistent with the first significant reduction in usage occurring around 9:00am and persisting.[23]

Another explanation is that the information treatments had an independent effect on daytime consumption that more than offsets the consumption-encouraging effects of lower daytime prices. However, this explanation is inconsistent with the placebo test presented in the online appendix, which shows negligible effects on

[20] Formally, $W_{i,t}$ is defined as follows:

$$W_{i,t} = \begin{cases} 1, & \text{if } i \text{ is treated } \textbf{and } t \text{ is in treatment period} \\ 0, & \text{otherwise} \end{cases}$$

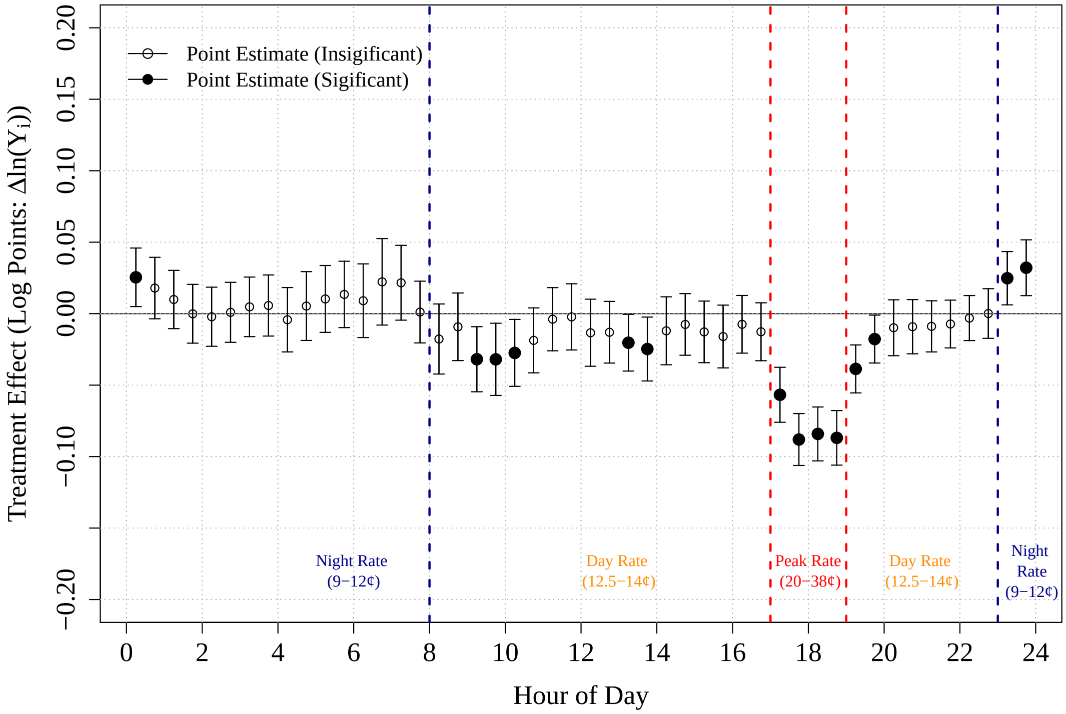
[21] I do not need to include treatment assignment and treatment period indicators as separate regressors because they are captured by the household and month-of-sample fixed effects, respectively.

[22] Unless otherwise stated, throughout the paper I am addressing local average treatment effects. As previously mentioned, compliance was imperfect due to attrition, meaning all my estimates must be considered LATEs. However, Commission for Energy Regulation (2011b) suggests attrition was largely unrelated to treatment status.

[23] A similar effect was found by Blonz (2016) for commercial and industrial users.

weekends and holidays, when TOU pricing was not in effect but the information treatments would still be salient.

A final explanation is that consumers respond to hour-to-hour *changes* in prices, rather than price levels. For example, consumption falls at 8am when prices increase from the night rate (very low) to the day rate (somewhat low). Similarly, consumption increases at 11pm when prices decrease from the day rate to night rate.



Notes: Error bars represent 95% confidence intervals. Solid dots are statistically significant at the 95% level; empty dots are not. The horizontal placement of points indicates the middle of time period (e.g., the first point, located at the x-coordinate of approximately 0.25, is the half hour spanning 12:00am – 12:30am, and the last point lying between the two red vertical lines spanning 18:30 – 19:00 (6:30pm – 7:00pm), which is the last half-hour of the peak period). Only days with active time-of-use pricing (non-holiday weekdays) are included. For each point estimate, N = 1,117,911. Standard errors are two-way clustered at the household and month-of-sample levels.

Figure 2: Average Treatment Effects by Time of Day

A simple way to observe what is driving the results is to consider the simple average daily consumption profiles by treatment and control groups, both during the baseline period and during the treatment period. This is shown in Figure 3. Two important features of this figure stand out. First, the consumption profile for the control group does not significantly change from baseline to treatment period. This suggests that the treatment period (January to December 2010) featured nearly identical average demand conditions as the baseline period (July to December 2009).[24] This reveals that the identification of the treatment effect is driven

[24] Note further that the consumption profiles look very similar, even though the baseline and treatment periods did not cover the same calendar year. The results are robust to only using the second half of the treatment period in order to align the calendar months across periods.

primarily by within-household changes in consumption patterns before and after the treatment period. In particular, a single difference of baseline and treatment period consumption for the treatment group (rather than difference-in-difference) yields almost exactly the same average treatment effect.

The lack of any strong time trend for the control group supports the parallel trends assumption required for the difference-in-differences estimator. For the parallel trend assumption to be violated, the treatment group would have had to exhibit a particular time trend absent treatment, even though the control group evidently did not. In the online appendix, I further assess pre-trends during the baseline period, finding that they are similar for the treatment and control groups. The other important assumption for difference-in-differences is the stable unit treatment value assumption (SUTVA), which requires that control households are not affected by treatment households (or visa versa). This is plausible because this was a nationwide program of only a few thousand geographically-dispersed households, so participating households are unlikely to have interacted on any significant scale.

The second evident feature is that treatment households do not ramp up their demand as rapidly during the treatment period as they did during the baseline period. This is the primary source of identification for the average treatment effect. For households assigned to treatment, consumption peaked at around 0.46 kWh[25] during the baseline period compared to about 0.407 kWh during the treatment period, a reduction of approximately 11%.[26]

As previously mentioned, the treatment group appears to have somewhat higher baseline consumption in peak periods than the control group does. This difference is not quite statistically significant at standard levels ($p = 0.07$). However, the difference is non-trivial in magnitude. As a result, a naive difference between the treatment and control groups would understate the treatment effect during peak hours by a factor of two, and it would estimate the wrong sign for the treatment effect during daytime hours. This highlights the importance of a difference-in-differences approach.

5 Heterogeneity in Treatment Effects

As shown in the previous section, the program led to substantial treatment effects during peak hours, and there is some evidence for smaller effects during off-peak hours. In this section, I analyze the heterogeneity (on observables and across treatment groups) in the treatment effects.

[25] All kWh figures are in units of energy consumed during the half hour interval. To convert to load in kW, multiply by 2. E.g., 0.46 kWh consumed during 30 minutes is equivalent to a rate of 0.92 kWh consumed per hour, or a load of 0.92 kW.

[26] This 11% reduction is slightly larger than the treatment effects shown in Figure 2 for four reasons: rounding, it is a difference rather than a difference-in-difference, the log approximation in equation (1), and Jensen's inequality (that is, the average of the log is not the same as the log of the average).

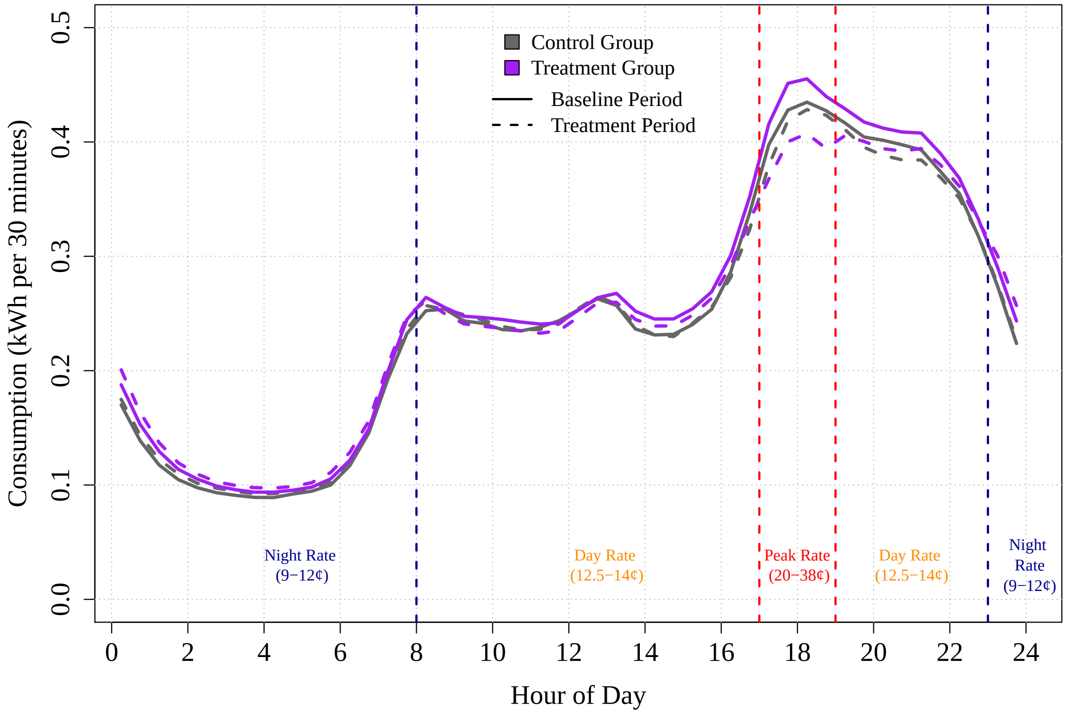


Figure 3: Average Consumption by Time of Day, Treatment Assignment, and Period

5.1 Method

5.1.1 Overview of Athey-Imbens Causal Tree Algorithm

I apply and extend the Athey and Imbens (2016) causal tree (CT) algorithm to estimate heterogeneous treatment effects. The algorithm uses cross-validated regression trees to estimate conditional average treatment effects (CATEs). CATEs represent the expected treatment effect, conditional on observable characteristics X_i. Using the standard potential outcomes notation (e.g., $Y_i(1)$ and $Y_i(0)$ for treated and untreated outcomes for household i), the CATE is denoted

$$\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]. \tag{2}$$

An example of a conditional average treatment effect is the average effect among college-educated women over the age of 65. With many different observables X_i, some of which may be continuous or interact with each other, estimating $\tau(x)$ is a daunting task, particularly if one wishes to estimate it non-parametrically.

The CT algorithm uses regression trees to estimate $\tau(x)$ in a parsimonious yet non-parametric manner.[27] In particular, the CT algorithm recursively splits the data along covariates into hyper-rectangles, estimating the treatment effect separately within each subset, called "leaves." This is called growing the tree. Setting

[27] This paper presumes that the reader has some understanding of regression trees.

aside how the tree is grown, once the splits are found, estimating the CATEs ($\tau(x)$) is straightforward. Given a tree (denoted Π), the CATE for an individual with characteristics x is simply the difference in conditional means for treatment and control outcomes falling in the same leaf as x (denoted $\ell(x, \Pi)$):

$$\hat{\tau}(x) \equiv \hat{\mu}(x; S_T, \Pi) - \hat{\mu}(x; S_C, \Pi), \tag{3}$$

where $\hat{\mu}(x; S, \Pi)$ is simply the conditional mean of observed outcomes in group $\mathcal{S}$ (i.e., treatment or control) in the leaf of tree Π where x falls.

$$\hat{\mu}(x; S, \Pi) \equiv \frac{1}{|\{i \in \mathcal{S} : X_i \in \ell(x, \Pi)\}|} \sum_{i \in \mathcal{S}: X_i \in \ell(x, \Pi)} Y_i. \tag{4}$$

This is analogous to simply computing the LATE for a each subgroup in the data, where the tree determines the subgroups.

The key innovations in Athey and Imbens (2016) are that it solves two practical difficulties in growing the tree. The first is how to grow the tree to begin with, which is far simpler for standard regression trees that are meant to predict a single outcome variable, Y_i, than for treatment effects τ. The second is how to determine the optimal size of the tree, since cross-validation techniques require the researcher to know the ground truth in the outcome of interest. In standard regression problems, the outcome of interest is observed (the outcome, Y_i), but in causal inference it is not (the treatment effect, τ). I apply the Athey and Imbens (2016) causal tree algorithm to my setting, while extending it to a difference-in-differences framework with multiple treatment groups, discussed next.[28]

5.1.2 Extensions to the Athey-Imbens Causal Tree Algorithm

The CT algorithm is designed for standard experiments with a single treatment group and a single observation per treated unit. I extend it to suit the circumstances of this data, which features multiple treatment groups and calls for a difference-in-differences strategy. These circumstances are very common, and so my extensions are generally applicable.

First, I extend this algorithm to a difference-in-differences framework where households are observed both before and after treatment. This extension is straightforward, as one need only replace the outcome variable Y_i in the CT algorithm with its change from pre- to post-treatment means, $\Delta Y_i \equiv \overline{Y}_{i,t'} - \overline{Y}_{i,t}$, where the mean is taken over time for each period, t' (the treatment period) and t (baseline period). This captures the first

[28] An even more sophisticated algorithm is causal forests, which further extends causal trees to use random forests. As a sensitivity, I also conducted this analysis. The results were very similar. I use the simpler causal tree but causal forests are much more difficult to interpret, as that method produces thousands of individual treatment effects that are not easily summarized in a table or graphic.

difference, and indeed has better statistical properties than panel data methods, as documented in Bertrand, Duflo, and Mullainathan (2004). The CT algorithm then computes the difference in these differences.[29]

Second, I extend the CT algorithm to estimate heterogeneity in treatment effects across multiple treatment groups. This extension is more substantial. The natural way to implement multiple treatment effects in the CT algorithm would be to include an indicator variable for each of the $m \in \{1, ..., M\}$ treatment groups, denoted W_i^m, allowing the algorithm to find heterogeneity in treatment effects across those indicators. The practical problem with this approach is that a split on any of these variables would result in no control observations in that side of the branch, making it impossible to estimate treatment effects.

The solution is to replace each control observation with M copies of it, each pseudo-assigned to a different treatment group. That is, for each control observation i and treatment $m \in \{1, ..., M\}$, generate a new pseudo-observation i_m with all observables the same ($Y_{i_m} \equiv Y_i$, $X_{i_m} \equiv X_i$, $W_{i_m} \equiv W_i$ $(=0)$) and a new set of variables:

$$\text{For } m' \in \{1, ..., M\} \quad W_{i_m}^{m'} \equiv \begin{cases} 1 & \text{for } m' = m \\ 0 & \text{for } m' \neq m, \end{cases}$$

for a total of $M \times n_C$ control observations, in place of the original n_C observations. This leads to a new pseudo-sample size of $\tilde{n} = (n_T + Mn_C)$. With this transformed dataset, the tree algorithm can split on the treatment group covariates (W_i^m) while continuing to use W_i as the treatment indicator. Meanwhile, duplicating control group observations does not alter the means of control outcomes or observables. As a result, the LATE and CATE estimates on the transformed data are numerically equivalent at every branch to the corresponding estimates on the untransformed data.

To see this formally, consider the mother node, before any splits have occurred, where there is no difference between the LATE and the CATE. Simply considering the LATE, it is easy to show that the LATE of the transformed data is numerically equivalent to the LATE of the untransformed data. The estimated LATE on the transformed dataset is (denoting the set of the transformed control observations $\tilde{\mathcal{S}}_C$)

$$\begin{aligned} \frac{1}{n_T} \sum_{i \in \mathcal{S}_T} Y_i - \frac{1}{M \times n_C} \sum_{i \in \tilde{\mathcal{S}}_C} Y_i &= \frac{1}{n_T} \sum_{i \in \mathcal{S}_T} Y_i - \frac{1}{M \times n_C} M \sum_{i \in \mathcal{S}_C} Y_i \\ &= \frac{1}{n_T} \sum_{i \in \mathcal{S}_T} Y_i - \frac{1}{n_C} \sum_{i \in \mathcal{S}_C} Y_i \\ &= \hat{\tau}, \end{aligned} \tag{5}$$

[29] If households are observed the same number of times before and after the treatment, the resulting estimates of the LATE and CATEs are numerically equivalent to the standard difference-in-differences estimates.

which is the same as the LATE estimate of the untransformed data.[30]

Moreover, the CATE of the transformed data at any node of the tree is also equal to the CATE for the original dataset, even after splitting on any combination of observables and/or treatment groups.[31]

When there are overlapping dimensions of treatment, the process can be simplified somewhat to reduce computation time. One need not create an additional observation and indicator variable for each permutation of the different treatment dimensions. For example, in my data, households received one of four information treatments and one of four pricing schedules, resulting in 16 distinct treatment groups. In this case, the process outlined above can be executed twice, with $M = 4$ each time, leading to 8 ($= 4 + 4$) copies of the control observations, rather than 16 ($= 4 \times 4$), resulting in a pseudo-sample size of $\tilde{n} = (n_T + (4 + 4)n_c)$.

5.2 Causal Tree Estimation Results

5.2.1 Results

I estimate the causal tree algorithm using difference-in-differences with multiple treatment groups as described in the previous section. In practice, this means that the outcome variable for household i is the percentage change in i's average peak consumption between the baseline and treatment period. In particular, this can be interpreted as estimating an equation of the following form:

$$\Delta Y_i = \mu + \tau W_i + \varepsilon_i \tag{6}$$

where ΔY_i is percentage change in household i's average peak-period consumption between the baseline and treatment period, and as before W_i is an indicator variable for treatment. This is a difference-in-differences estimate because ΔY_i is the first difference (before/after for each household), and τ represents the mean difference in these differences between the treatment and control groups.[32] Hence, τ is the difference-in-

[30] While the point estimates of the transformed data are equivalent to those of the original data, the duplicated observations are perfectly correlated (i.e., the observations are not independent) and threaten to bias standard errors. To correct for this, one can either cluster standard errors at the household level or compute standard errors using using the original dataset with standard methods. A related issue arises in cross validation because now households appear in the data multiple times, and standard resampling methods are likely to assign a single household's multiple pseudo-observations to multiple CV groups. To solve this, households should be block-assigned to CV groups.

[31] Showing this involves only two differences compared to the logic in equation (5). First, the means are conditional on X_i and W_i^m. And second, the values of M in the second term are decremented by one for each treatment group that has previously been split upon, as those groups have been previously diverted into another branch. This latter difference does not affect the estimated CATE, because the decrement appears in both the numerator and the denominator of the second term, and therefore cancels.

[32] This differs from the panel-data-style difference-in-differences estimator, which involves method involves regressing $Y_{i,t}$ (as opposed to its change over time) on indicators representing both treatment group and treatment period. In order to apply the causal tree algorithm, I must difference baseline and treatment consumption to obtain a single outcome per household. This is because the causal tree algorithm is designed for cross-sectional data, not panel data. Estimating the treatment effects using the panel method produces nearly identical estimates, differing only slightly due to the log approximation error and the different number of hours in the baseline and treatment periods. While one may think that using the larger sample size of a panel dataset would produce smaller standard errors, in these data the resulting standard errors are approximately the same in both methods after clustering. This is because clustering accounts for the very high within-household correlation in energy consumption over time. See Bertrand, Duflo, and Mullainathan (2004) for evidence that aggregating extended time-series into a small number of observations (as I do here) can actually improve the accuracy of standard errors in difference-in-differences estimators.

differences estimator for the average treatment effect. The causal tree algorithm determines the subgroups of the data for which to estimate the τ separately, giving the treatment effect as a function of covariates, or $\tau(x)$.

Figure 4 shows the results from the causal tree estimation for peak consumption. Each box in that figure represents a "node" of the tree, corresponding to the particular subset of the samples satisfying all conditions from the branches above it. Each box shows the estimated treatment effect (TE), its standard error (SE), and the number of treated observations in that node.[33] The colors of the nodes correspond to the size of the treatment, with green corresponding to larger reductions in peak consumption.

The top node (node [1]) shows the local average treatment effect (LATE) for the full dataset of -8.9% of baseline peak consumption. The first split in the data found by the CT algorithm is awareness: households that reported being aware of the change in their tariff structure exhibit a 4.5-times larger response (-10.3%, node [2]) compared those who were not (-2.3%, node [11], not statistically significant).

Beyond awareness, the treatment effect is estimated to be larger for households with higher baseline consumption, both in percentages and in levels. Among aware households, the estimated effect was -11.2% (node [3]) for households with baseline consumption of over 0.12 kWh per half-hour, compared to a noisy estimate of +8.4% (node [10]) for those below. This threshold of 0.12 kWh is very low, at the 5th percentile of average baseline peak consumption. Baseline consumption also matters for unaware households, with households above the 0.25 kWh threshold (the 24th percentile) reducing by -4.2% (node [12]) compared to -0.2% (node [13], not statistically significant) for those below it.[34]

Among aware households with more than 0.12 kWh baseline consumption, the remaining sources of heterogeneity stem entirely from the information treatments. The in-home electricity display (IHD) produced the strongest effect on peak consumption (-14.6%, node [4]), followed by the monthly bill (-11.7%, node [6]), then by the bi-monthly bill with the overall load reduction (OLR) incentive (-10.7%, node [8]), then by the bi-monthly bill treatment (-7.9%, node [9]).

These differences in effects are likely due to the information treatments amplifying the effectiveness of time-of-use pricing, rather than having independent effects on consumption. This is because there is little effect on weekends and holidays, when time-of-use pricing was not applied but the information treatment remained.[35]

No other features of the data proved to be robustly related to treatment effect heterogeneity.[36] This suggests that, conditional on these factors (awareness, baseline consumption, and information provision), no other observables are robustly related to heterogeneity in treatment effects.

[33] The standard errors are computed using an OLS regression of equation (6) for each relevant subgroup. Note that p-values based on these standard errors will be biased because the model was chosen using this data. The method for computing p-values after model selection methods such as this is not yet settled in the literature.

[34] One concern is that such differences are spurious, in that larger users would naturally exhibit larger changes in levels. However, all treatment effects here in are percentage changes, implying that the effect is not limited to changes in levels.

[35] See online appendix for the treatment effects during weekends and holidays, when time-of-use pricing was not in effect.

[36] While the first stage of the CT algorithm initially estimates a larger tree with richer heterogeneity than shown here, the second stage finds that those additional splits are not found to be robust to cross-validation. The tree shown in Figure 4 is optimally-sized tree by 10-fold cross-validation

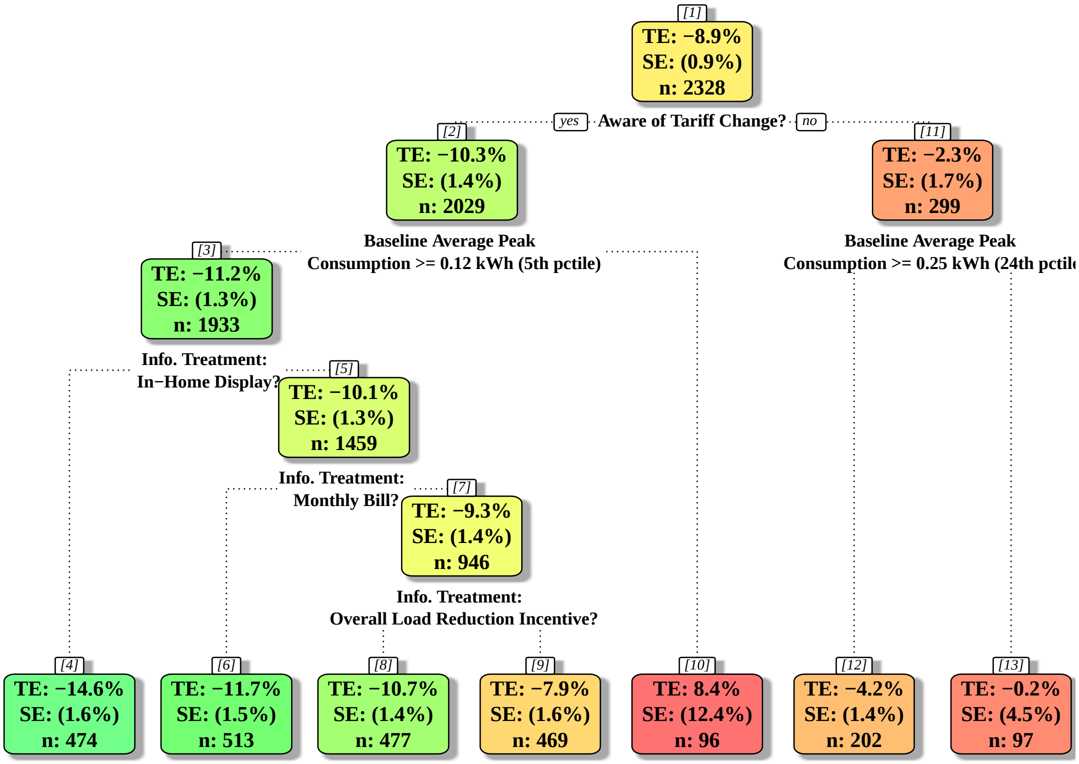
I also estimate treatment effects during nighttime and off-peak daytime hours, during which prices were somewhat lower for the treatment group. The program increased nighttime consumption by +1.8% (with standard error of 1.1%) and decreased daytime consumption by -2.2% (with a standard error of 0.7%). The algorithm finds no heterogeneity in these treatment effects.

5.2.2 Implications

These results suggest that the key to designing effective time-of-use pricing policy lies in information and awareness, with some role for targeting households with non-trivial baseline consumption. Other than baseline consumption, the only dimensions that appear to matter are the awareness of the pricing policy and the amount of information provided to households. Conditional on these factors, no other source of heterogeneity is robust to cross-validation methods.

This suggests that complicated household targeting mechanisms are not necessary to produce good results. Conditional on awareness and information provision, baseline consumption appears to be a sufficient statistic for any other observables that may affect the impacts of time-of-use pricing.

In addition, the model finds no heterogeneity along tariff structure level, the other key dimension of treatment that the trial was designed to test. I turn to this in the next section.



Notes: TE = Treatment effect; SE = standard error; n = number of treated observations in node. If the condition is satisfied, proceed down left branch. This figure was generated using the `rpart.plot` package (Milborrow 2016) in R.

Figure 4: Heterogeneous Treatment Effects on Consumption during Peak Periods

5.3 Treatment Effects by Tariff Levels

5.3.1 Estimated Demand Curves

The results above found no robust heterogeneity in effects other than awareness, information provision, and baseline consumption. This implies that the higher prices did not lead to significantly larger reductions, which may be surprising to economists. In this section, I explicitly investigate the effects of different tariff levels to confirm this previous finding.

In particular, I use the same difference-in-differences estimator from the previous section to estimate treatment effects by tariff group for peak, night, and daytime offpeak periods separately. I use these results to draw an average demand curve. This is possible because of the large variation in prices that households faced during peak periods. The results are shown in 5.

The demand curve during peak periods, shown on the right-hand side of the figure, is very inelastic at high prices. The smallest price increase (from the control of 14.1 to 20 € cents per kWh) features a significant reduction, but further increases yield strongly diminishing returns. While the lowest tariff treatment (20 € cents) resulted in a statistically smaller effect than the higher tariff treatments did ($p = 0.013$),[37] an F-test fails to reject the null that all four tariffs have jointly equal effects on peak consumption ($p = 0.08$). This is particularly surprising because, under the commonly-used assumption of a constant elasticity, reductions should scale proportionally with the magnitude of the price increase, but I find no evidence that it increases at all.

This is not simply a result of weak statistical power, as the size of the standard errors rule out proportional increases. For example, the second peak tier (€26 cents, and increase of €11.9 cents) represents approximately twice as large of a price increase of the first tier (€20 cents, an increase of €5.9 cents). Given a constant elasticity, the second tier would be expected to lead to twice as large a reduction. An effect of this magnitude would be easily detectable at the 5% significance level given the estimated standard errors. In particular, any difference between these two tiers in excess of 40% would be statistically detectable at the 5% level, and that would nonetheless still be consistent with a declining elasticity.[38]

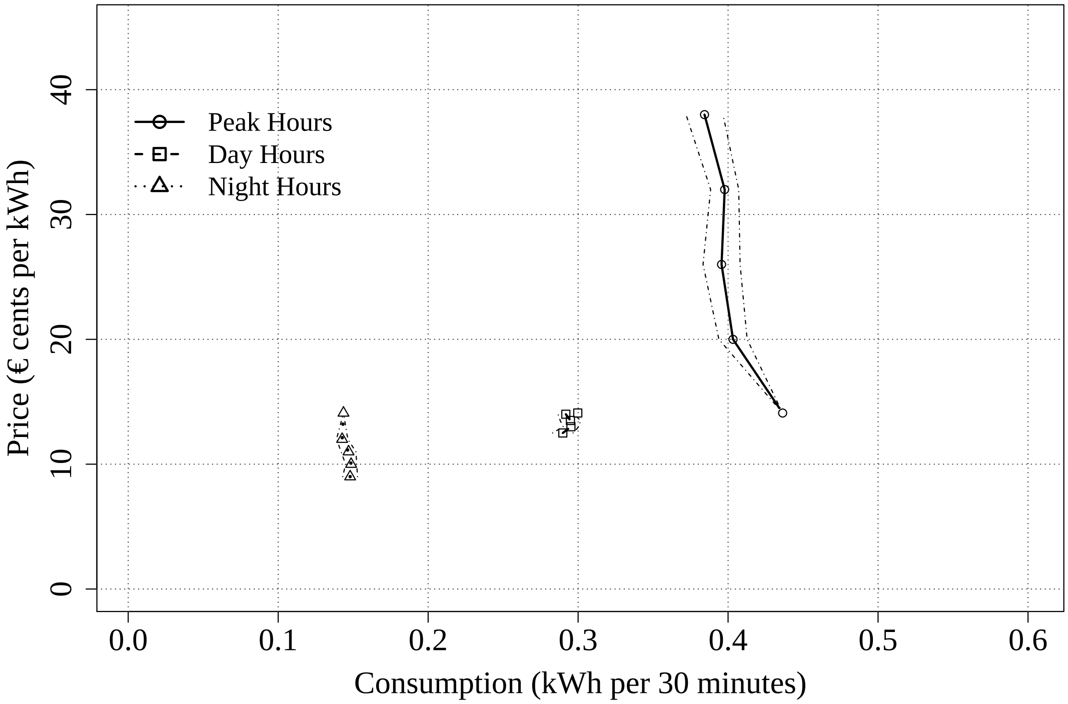
The demand curve during daytime (off-peak) periods is in the middle of the figure. There is little price variation during off-peak periods, but the effects during this period have a counter-intuitive sign, with consumption decreasing despite lower prices. This was previously discussed in section 4 and is likely due to households preemptively turning off devices in anticipation of higher prices during peak periods. Here, it manifests as an apparently upward-sloping demand curve to contemporaneous prices.

[37] This result was already found by Commission for Energy Regulation (2011b).

[38] In this data, the point estimate for the second tier is only 25% larger than the that of the first tier, well within the confidence interval.

During the night, there is a strongly inelastic demand curve, which can be seen in the lower left-hand side of the figure.

Altogether, it appears that the effects of time-of-use pricing depend on its mere existence, and not the level of the prices. This confirms the results from the previous section, which found that the causal tree algorithm found no apparent heterogeneity in the treatment effects by price level. There are several interpretations for this result. One interpretation is that households face strongly convex costs of reducing peak energy use, with small reductions being easy to achieve (turning off lights) but larger ones being very costly (turning off major appliances). An alternative interpretation is that households respond to the knowledge of differential pricing, but are not attentive to the specific level of the price.[39]



Notes: Each line represents estimated demand curves by period of the day (night, off-peak daytime, and peak). To construct these curves, I estimated treatment effects in percentage changes using a difference-in-differences (as in the causal tree estimates) for each tariff group and time period separately. I then transformed these percentage changes to average demand curves by multiplying the percentage change at each price level by the average baseline consumption in the relevant periods. Hence, these should be thought of as demand curves for the average consumer. 95% confidence intervals for the mean effects are shown, representing the uncertainty about the average treatment effects relative to the control group, which faced prices of 14.1 € cents.

Figure 5: Estimated Demand Curves, by Pricing Period

[39] An important caveat to this result is that it derives from the behavior of Irish residential consumers, and other types of consumers may respond differently. In particular, the households did not receive automation technology, such as programmable thermostats. In addition, Irish electricity demand does not primarily derive from easily-automated home heating and cooling demand. (Home heating in Ireland is typically fueled by oil or gas. Air conditioning relatively uncommon in Ireland due to its climate. Water heating is typically fueled by electricity, but on-demand heating is common in Ireland, which is not as easily automated as home heating.) Newer technology allows users set their thermostats to automatically adjust their set points in response to real-time prices, which may induce more sensitivity to price levels. Therefore, these results suggest that the price level is not a strong driver of peak demand *in the absence of automation technology.*

5.3.2 Implications of Declining Elasticities for Real-Time Pricing

Regardless of the reason for this declining elasticity, it suggests that real-time pricing is unlikely to gain much efficiency benefit over TOU pricing in contexts similar to Irish households. If the exact price level does not matter—only that peak usage is priced differently—then real-time pricing is unlikely to change hour-by-hour electricity consumption. If real-time pricing does not affect behavior, then it will not reduce peak load when needed, raising questions about whether it offers any real efficiency benefit.

To illustrate this point, I run a simulation of the effects of real-time pricing on consumption. I first estimate demand curves shown in Figure 5 separately for each "leaf" sub-group identified by the causal tree (Figure 4). These heterogeneous demand curves generally show similarly small and declining elasticities.[40] Using these demand curves, I simulate individual household consumption under three different kinds of pricing: flat pricing (14.1 cents per kWh), TOU pricing as implemented in this experiment, and real-time wholesale spot electricity prices in Ireland.[41]

Figure 6 shows total consumption across all simulated households for a day with particularly high demand and spot prices: January 4, 2010. Focusing on the peak periods, we see that TOU pricing generally reduces consumption relative to flat pricing (green versus blue lines) across the entire pricing window, including the hour of highest consumption. This indicates that TOU pricing successfully reduces peak load.

By contrast, real-time pricing does not reduce peak load and may even increase it. Wholesale electricity prices spiked that day, and the simulation suggests that this would have reduced consumption during the high-priced period. However, the declining elasticity implies that the response to this price spike would be very similar to the response to standard TOU pricing. Further, while real-time prices did indeed spike on this peak-load day, they did not spike during the peak hour. Large price spikes such as these are often fleeting because lower-cost suppliers cannot instantaneously ramp up to meet demand, but can do so on relatively short notice. In other words, because of adjustments on the supply side, the period of peak demand is not always the same as the period of peak prices. In this example, during the period of peak demand, the market price was *lower* than the flat retail price. As a result, households would be charged lower prices during this period under real time pricing than they would under a flat tariff, resulting in even higher peak load. This shows how real-time pricing can actually be strictly worse than flat-rate pricing from the perspective of peak load.

[40] The exception is the group in node [10] which showed a statistically insignificant increase in consumption in response to higher prices, suggesting an upward-sloping demand curve. Because of the counter-intuitive sign for this small subgroup, I treat this group as unresponsive for this simulation. I also treat daytime consumption as unresponsive, due to its similarly counter-intuitive sign. Because the focus of this simulation is peak consumption, this choice is largely immaterial.

[41] Spot prices represent "EP2" final prices collected from `http://www.sem-o.com/marketdata/Pages/dynamicreports.aspx`. I use the shadow prices (without capacity uplifts) because that represents true marginal generation costs, but including the uplift has very little effect on this simulation. Average spot prices are lower than average retail prices because they do not include transmission and distribution charges. I effectively assume that those costs would be recovered through a fixed charge on households' bills.

Of course, Figure 6 assumes that households would respond instantaneously to fast-changing prices under a real-time pricing regime. This is clearly a strong assumption, but the alternative would make real-time pricing appear even less effective. In particular, a more realistic assumption is that households respond less to real-time pricing than to predictable TOU pricing. TOU prices are set and known in advance. Real-time prices change every half hour and so would require households to pay constant attention to their meters, which would be very costly and infeasible for many households. If households would not respond instantaneously, the impact of hour-to-hour changes in prices would be more muted than TOU prices would be, making real-time pricing even less effective at affecting behavior than the simulation suggests.

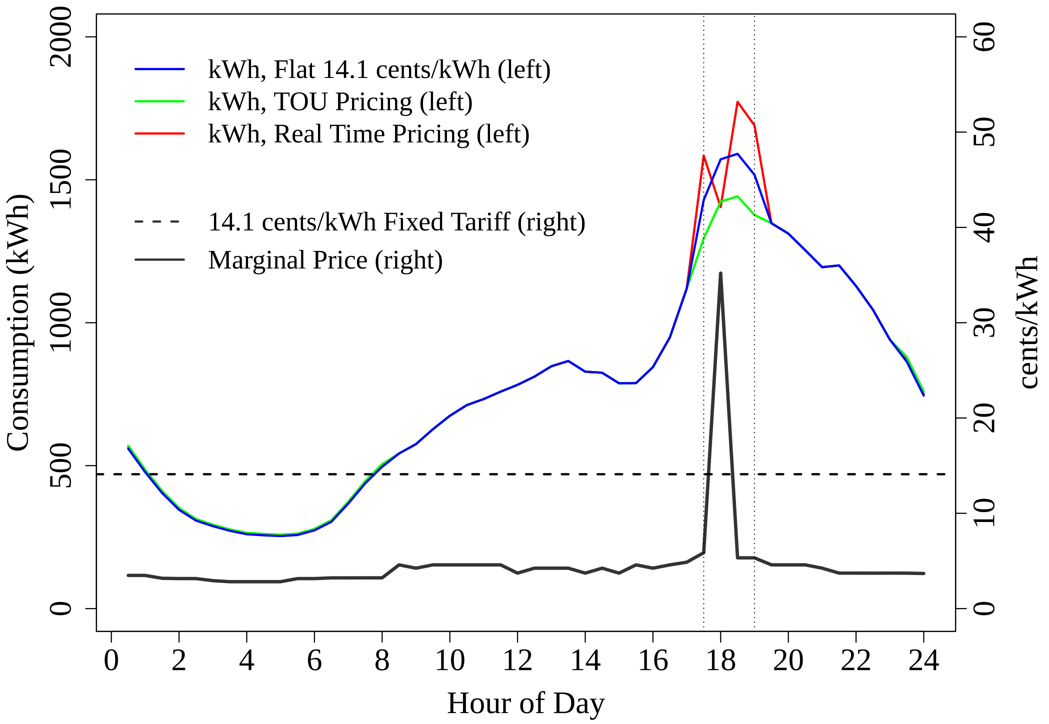


Figure 6: Simulated Consumption under Different Pricing Regimes: January 4, 2010

5.4 Who Is Likely to Be Aware?

Since awareness of time-of-use pricing is key to its effectiveness, is it possible to identify *ex ante* which households are likely to understand the policy? If so, policy can be targeted towards those households to improve outcomes. I answer this question using three different methods, all of which reach the same conclusion: household observables have only small predictive power for awareness.

The first method is involves estimating a linear probability model of awareness as a function of pre-treatment household observables and treatment group indicators.[42] The truncated set of the results is shown in column

[42] Only treatment households are included in this analysis, as they are the relevant group. A logistic model produces very similar results for this entire section, but its coefficients are more difficult for the reader to interpret. The typical disadvantage of linear

(1) Table 4 (all significant variables are shown, and the other 98 variables are suppressed on account of space). Of the 123 variables considered, 14 (11%) are significant at the standard 5% level. This is somewhat larger than the 5% expected if there were no explanatory power of the observables, but the low share of significant coefficients suggests caution in over-interpreting the ones that do turn out significant. In addition, no coefficients are significant with a Bonferroni correction.

To determine how many of these significant coefficients are spurious due to sampling variability, I employ a machine learning method, post-selection Lasso regression, to eliminate spurious variables from the model. This involves estimating the linear probability model with a penalty term for large and non-zero coefficient estimates to determine which variables to retain in the model.[43] The magnitude of the penalty parameter is chosen through cross-validation methods. I present results using two standard criteria for the optimal penalty parameter. Column (2) of Table 4 shows the results using the penalty parameter that minimizes the out-of-sample root mean square error (RMSE) according to cross validation.

Another standard way to choose the Lasso penalty parameter is to use the most parsimonious model (i.e., the largest penalty parameter) that produces an out-of-sample RMSE that is within one standard error of the minimum RMSE.[44]. This penalty parameter results in the much simpler model shown in column (3). This model drops all variables except for an indicator for whether the household has internet access in the home. Those with internet access are 13 percentage points more likely to be aware of their pricing regime (91% versus 78%). While a 13 percentage point difference is non-trivial, it is somewhat modest given that it is the most important predictor according to the Lasso results.[45] Altogether, this suggests that there is not a clearly reliable way to predict whether households will be aware of the TOU pricing *ex ante*.

To further illustrate this weak power that observables have to predict awareness, Figure 7 shows histograms of fitted probabilities using each model in Table 4.[46] These are the probabilities one would want to compute

probability models—that they may imply predicted probabilities outside of the interval $[0, 1]$ is not a significant problem in this case, as the vast majority of predicted probabilities fall within that interval.

[43] Lasso regressions are estimated by minimizing a standard loss function (such as the sum of the sum of square residuals or negative log likelihood) *plus* a penalty term proportional to the L1-norm ($|| \cdot ||$) of the standardized coefficient vector (here denoted β). In the context of a linear model, the objective function is

$$\min_{\beta} \sum_{i} (y_i - X_i'\beta) + \lambda ||\beta||$$

where all X_i variables have been standardized to have a mean of zero and a standard deviation of one. Depending on the value of λ, the results range from collapsing to exactly the OLS estimates (when $\lambda = 0$) to a null model with no covariates (for sufficiently large λ). The optimal value of λ is chosen by estimating the out-of-sample prediction error through cross-validation. This process completely eliminates some variables of the model, while retaining ones that have out-of-sample predictive power.

I use post-selection Lasso, which means estimating a Lasso model, determining which variables are retained, then re-estimating the model by OLS using only those retained variables.

[44] This is motivated by the fact that the model produces a statistically indistinguishable error rate, but is more parsimonious and so avoids overfitting. See James, Witten, Hastie, and Tibshirani (2013).

[45] Regression trees also find that internet access is the most important predictor for awareness, but cross validation excludes it as non-robust.

[46] This figure truncates fitted probabilities from the linear model that are greater than 100%, replacing them with 100%. Results from a logistic regression, which do not feature fitted probabilities exceeding 1, look similar. In addition, histograms of the fitted probabilities from a much more flexible random forest model are also similar, implying that these results are not the result of an insufficiently flexible functional form.

if policymakers wanted to target TOU pricing on observables, perhaps by preferentially treating those who are more likely to be aware. In general, they show that few households exhibit low probabilities of awareness, suggesting that ignoring such households in a policy roll-out would have little effect. In the full model, the lowest predicted probability of awareness is about 40%, and only 0.4% of households are more likely to be unaware than aware.[47] In the RMSE-minimizing Lasso model, the lowest predicted probability of awareness is about 60%, implying that only targeting households that are more likely to be aware than unaware would have no effect on the treatment. The simplest model (the Lasso using the 1 standard error rule for the penalty parameter) leads to only two predicted probabilities: 91% for those with internet access, and 78% for those without. All of these results suggest little role for using observables to target TOU pricing towards households likely to be aware.

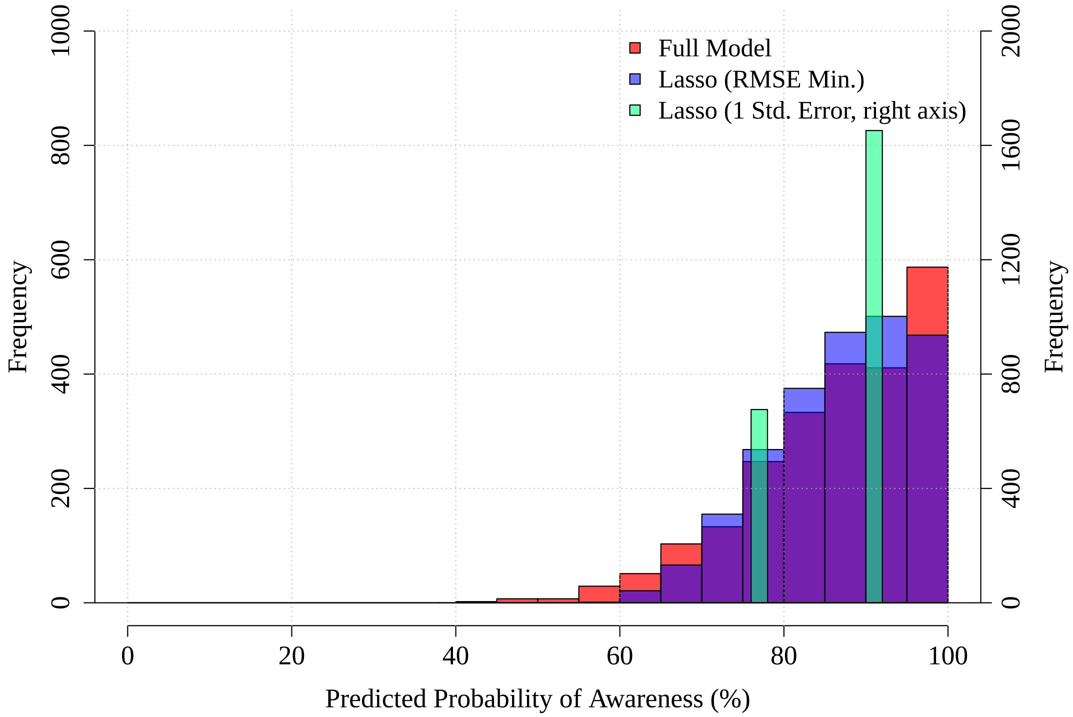


Figure 7: Histograms of Fitted Probabilities of Awareness, from Models shown in Table 4

6 Conclusion

I apply and extend new machine learning methods to estimate heterogeneous treatment effects from time-of-use pricing and information provision on residential electricity consumption. Most importantly, the effect of time-of-use pricing on peak energy consumption is 4.5 times larger for households who are aware of the change

[47] That is, in the full model, only 0.4% of households have a probability of awareness under 50% according to their observables.

Table 4: Post-Selection Lasso Linear Probability Model of Awareness

	Dependent variable:		
	Aware of Tariff Change (Indicator)		
	(1)	(2)	(3)
Internet Access in Home (Indicator)	0.06	0.06	0.13
	(0.02)	(0.02)	(0.02)
Use Internet Regularly (Indicator)	0.02	0.01	
	(0.01)	(0.01)	
Number of Children under 15 in Home	0.01	0.01	
	(0.01)	(0.01)	
Already Made Changes to Reduce Electricity Use (Pre-Trial)	0.01	0.01	
	(0.01)	(0.01)	
Heating Fuel: Oil (Indicator; “none” omitted)	−0.06		
	(0.03)		
Water Heating Fuel: Oil (Indicator; “none” omitted)	0.05	0.04	
	(0.02)	(0.01)	
Number of Dishwashers in Home	0.03	0.04	
	(0.02)	(0.02)	
Number of Desktop Computers in Home	0.02	0.01	
	(0.01)	(0.01)	
Share of Windows Double Glazed (1-5 = 0/25/50/75/100%)	0.01	0.01	
	(0.01)	(0.01)	
Expect Participating in Trial Will Reduce My Bill (Indicator)	0.05	0.06	
	(0.03)	(0.02)	
Expect to Choose More Efficient Appliances*	−0.03		
	(0.01)		
Satisfied with Share of Electricity From Renewables*	0.01	0.01	
	(0.01)	(0.01)	
Satisfied with Ability to Sell Extra Electricity to Grid*	0.01	0.01	
	(0.01)	(0.01)	
Female Respondent (Indicator)	0.04	0.04	
	(0.02)	(0.02)	
Social Class: AB (Highest) (Indicator, “Refused” omitted)	−0.02	0.02	
	(0.07)	(0.02)	
Education: Third (e.g., University) (Indicator, “Refused” omitted)	0.05	0.04	
	(0.03)	(0.02)	
Own home with mortgage (Indicator, “Rent from private owner” omitted)	0.06	0.03	
	(0.08)	(0.02)	
Cook stove type: Electric (Indicator; solid fuel omitted)	0.11		
	(0.05)		
Cook stove type: Gas (Indicator; solid fuel omitted)	0.09		
	(0.05)		
Info. Treatment: Monthly bill (Indicator; Bi-mon. omitted)	0.04	0.05	
	(0.02)	(0.02)	
Info. Treatment: In-Home Display (Indicator; Bi-mon. omitted)	0.05	0.05	
	(0.02)	(0.02)	
Info. Treatment: OLR (Indicator; Bi-mon. omitted)	−0.01		
	(0.02)		
Tariff Treatment: B (Indicator; A omitted)	−0.02		
	(0.02)		
Tariff Treatment: C (Indicator; A omitted)	−0.01		
	(0.02)		
Tariff Treatment: D (Indicator; A omitted)	0.02		
	(0.02)		
Constant	1.06	0.51	0.78
	(0.26)	(0.04)	(0.01)
Observations	2,328	2,328	2,328
R^2	0.12	0.07	0.03
Adjusted R^2	0.07	0.07	0.03
F Statistic	2.44	10.53	74.23
Number of Covariates	123	17	1
Number of Covariates Not Shown	98	*na*	*na*
Number of Covariates Significant (5% level)	14	*na*	*na*
Share of Covariates Significant (5% level)	11.4%	*na*	*na*

*p<0.1; **p<0.05; ***p<0.01

Notes: Due to space limitations, only a subset of the 123 included variables in column (1) are shown. The variables shown include all statistically significant ones at the 5% level, all variables retained by the Lasso technique, and all treatment group indicators. The full set of results is available upon request. Column (2) shows the post-Lasso selected model using the model that minimizes the out-of-sample root mean square error (RMSE). Column (3) shows the most parsimonious model that has an error within one standard error of the RMSE of the RMSE-minimizing model shown in (2). p-values and significance levels not presented for columns (2) and (3) because classical tests are not applicable the post-selection models. Methods for inference in post-selection models is a contentious area of on-going research. Baseline Average Consumption variables units are kWh per 30 minute interval. Approximately 70 percent of all households have internet in their home. In column (1), the excluded indicators for social class and education are “refused to answer”, and the excluded indicator for “own home with mortgage” is “rent from private owner”. In columns (2) and (3), the excluded indicators are all other possible outcomes.

* These variables featured numeric responses, where respondents reported on a 1-5 scale to what extent they agree (1) or disagree (5) with the statement.

This table was generated using the `stargazer` package (Hlavac 2015) for R.

in their pricing structure compared to those who are not (-10% versus -2.3%). Beyond awareness, the treatment effect varies significantly based on baseline electricity consumption and the amount of information provided to consumers about their dynamic pricing and energy usage. Households with very low baseline energy use do not reduce their consumption on average. Meanwhile, the strongest information treatment (an in-home electricity monitor) leads to nearly twice the reduction in peak consumption (-15%) compared to the weakest information treatment (a bi-monthly energy use statement, -8%).

According to cross-validation techniques, no other observables—or permutations thereof—are robustly related to treatment effect heterogeneity (conditional on the above factors of awareness, baseline consumption, and information treatment). This includes considering potential heterogeneity on more than 150 observables encompassing socio-demographic characteristics, attitudes towards energy and environmental issues, housing attributes, and household appliance characteristics. I also show that awareness is not reliably predictable even with rich information about observable household characteristics, which suggests that attempting to target TOU pricing based on awareness is unlikely to prove fruitful, although targeting on baseline consumption could be be effective. I also show that larger price increase do not induce larger responses, which suggests that the significant attention economists pay towards fine-tuning retail prices is less important than getting consumers to pay attention in the first place.

References

Athey, S., and G. Imbens (2016): "Recursive partitioning for heterogeneous causal effects," *Proceedings of the National Academy of Sciences*, 113(27), 7353–7360.

Athey, S., J. Tibshirani, and S. Wager (2016): "Solving Heterogeneous Estimating Equations with Gradient Forests," *ArXiv e-prints*.

Bertrand, M., E. Duflo, and S. Mullainathan (2004): "How much should we trust differences-in-differences estimates?," *The Quarterly journal of economics*, 119(1), 249–275.

Blonz, J. A. (2016): "Making the Best of the Second-Best: Welfare Consequences of Time-Varying Electricity Prices," .

Bollinger, B. K., and W. R. Hartmann (2016): "Welfare Effects of Home Automation Technology with Dynamic Pricing," *Working Paper*.

Burlig, F., C. Knittel, D. Rapson, M. Reguant, and C. Wolfram (2017): "Machine Learning from Schools about Energy Efficiency," Discussion paper, Working Paper.

Cappers, P., C. A. Spurlock, A. Todd, P. Baylis, M. Fowlie, and C. Wolfram (2017): "Time-of-Use as a Default Rate for Residential Customers: Issues and Insights," Discussion paper, Working Paper.

Carroll, J., S. Lyons, and E. Denny (2014): "Reducing household electricity demand through smart metering: The role of improved information about energy saving," *Energy Economics*, 45, 234–243.

Chetty, R., A. Looney, and K. Kroft (2009): "Salience and Taxation: Theory and Evidence," *American Economic Review*, 99(4), 1145–77.

Commission for Energy Regulation (2011a): "Appendices to Electricity Smart Metering Customer Behaviour Trials (CBT) Findings Report," *Appendices to Information Paper*.

——— (2011b): "Electricity Smart Metering Customer Behaviour Trials (CBT) Findings Report," *Information Paper*.

Deng, A., P. Zhang, S. Chen, D. W. Kim, and J. Lu (2016): "Concise Summarization of Heterogeneous Treatment Effect Using Total Variation Regularized Regression," *ArXiv e-prints*.

Di Cosmo, V., D. O'Horaa, and N. Devitt (2015): "Nudging Electricity Consumption Using TOU Pricing and Feedback: Evidence from Irish Households," *ESRI Working Paper No. 508*.

Faruqui, A., and J. Palmer (2012): "The Discovery of Price Responsiveness–A Survey of Experiments Involving Dynamic Pricing of Electricity," *Available at SSRN 2020587*.

Finn, P., and C. Fitzpatrick (2014): "Demand side management of industrial electricity consumption: promoting the use of renewable energy through real-time pricing," *Applied Energy*, 113, 11–21.

Fripp, M. (2016): "Making an Optimal Plan for 100% Renewable Power in Hawaii - Preliminary Results from the SWITCH Power System Planning Model," Discussion paper.

Herter, K., and S. Wayland (2010): "Residential Response to Critical-Peak Pricing of Electricity: California Evidence," *Energy*, 35(4), 1561–1567.

Hlavac, M. (2015): *stargazer: Well-Formatted Regression and Summary Statistics Tables.*R package version 5.2.

Ito, K., T. Ida, and M. Tanaka (2015): "The Persistence of Moral Suasion and Economic Incentives: Field Experimental Evidence from Energy Demand," *Working Paper*.

Ivanov, C., L. Getachew, S. A. Fenrick, and B. Vittetoe (2013): "Enabling technologies and energy savings: The case of EnergyWise Smart Meter Pilot of Connexus Energy," *Utilities Policy*, 26, 76–84.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013): *An Introduction to Statistical Learning*, vol. 6. Springer.

Jessoe, K., and D. Rapson (2014): "Knowledge Is (Less) Power: Experimental Evidence from Residential Energy Use," *American Economic Review*, 104(4), 1417–38.

McCoy, D., and S. Lyons (2016): "Unintended outcomes of electricity smart-metering: trading-off consumption and investment behaviour," *Energy Efficiency*, pp. 1–20.

Milborrow, S. (2016): *rpart.plot: Plot rpart Models. An Enhanced Version of plot.rpart*R package.

Pon, S. (2015): "Effectiveness of Real Time Information Provision with Time of Use Pricing," *FCN Working Paper No. 8/2015*.

Thorsnes, P., J. Williams, and R. Lawson (2012): "Consumer Responses to Time Varying Prices for Electricity," *Energy Policy*, 49, 552–561.

Torriti, J. (2012): "Price-based demand side management: Assessing the impacts of time-of-use tariffs on residential electricity demand and peak shifting in Northern Italy," *Energy*, 44(1), 576–583.

Wolak, F. A. (2011): "Do Residential Customers Respond to Hourly Prices? Evidence from a Dynamic Pricing Experiment," *American Economic Review*, 101(3), 83–87.