UNIVERSITY OF CAMBRIDGE | Energy Policy Research Group

Causal Tree Estimation of Heterogeneous Household Response to Time-Of-Use Electricity Pricing Schemes

EPRG Working Paper 1906

Cambridge Working Paper in Economics 1865

Eoghan O'Neill and Melvyn Weeks

Abstract We examine the distributional effects of the introduction of Time-of-Use (TOU) electricity pricing schemes. Using a causal forest (Athey and Imbens, 2016; Wager and Athey, 2017), we consider the association between past consumption and survey variables, and the effect of TOU pricing on household electricity demand. We describe the heterogeneity in household variables across quartiles of estimated demand response and utilise variable importance measures. Given that a number of standard variable importance measures can be biased towards continuous variables, we include permutation-based tests for our variable importance results.

Keywords Machine learning, TOU tariffs, Smart metering, Household electricity demand.

JEL Classification Q41, C55.

Contact mw217@cam.ac.uk
Publication January 2019

Causal Tree Estimation of Heterogeneous Household Response to Time-Of-Use Electricity Pricing Schemes

Eoghan O'Neill
Faculty of Economics
University of Cambridge

Melvyn Weeks*
Faculty of Economics and Clare College,
University of Cambridge

January 18, 2019

Abstract

We examine the distributional effects of the introduction of Time-of-Use (TOU) pricing schemes where the price per kWh of electricity usage depends on the time of consumption. These pricing schemes are enabled by smart meters, which can regularly (i.e. half-hourly) record consumption. Using causal trees, and an aggregation of causal tree estimates known as a causal forest (Athey & Imbens 2016, Wager & Athey 2017), we consider the association between the effect of TOU pricing schemes on household electricity demand and a range of variables that are observable before the introduction of the new pricing schemes. Causal trees provide an interpretable description of heterogeneity, while causal forests can be used to obtain individual-specific estimates of treatment effects.

Given that policy makers are often interested in the factors underlying a given prediction, it is desirable to gain some insight to which variables in this large set are most often selected. A key challenge follows from that fact that partitions generated by tree-based methods are sensitive to subsampling, while the use of ensemble methods such as causal forests produce more stable, but less interpretable estimates. To address this problem we utilise variable importance measures to consider which variables are chosen most often by the causal forest algorithm. Given that a number of standard variable importance measures can be biased towards continuous variables, we address this issue by including permutation-based tests for our variable importance results.

JEL Classification Codes: Q41, C55.
Keywords: Machine learning, TOU tariffs, Smart metering, Household electricity demand.

*Contact Author: Dr. M. Weeks, Faculty of Economics, University of Cambridge, Cambridge CB3 9DD, UK. Email: mw217@econ.cam.ac.uk. Our thanks are due to Kai Liu, David Newbery, Alexei Onatski, and Michael Pollitt.

Contents

List of Tables

List of Figures

1 Introduction

If a policymaker believes the impact of a particular policy are heterogeneous in a given population, then it is helpful to report the distributional effects of the policy. The critical question is: does the policymaker know ex ante which characteristics of individuals are driving the differences in the impact of the policy?

Increasingly researchers have many available covariates at their disposal and it may not be clear which covariates should be used to categorise heterogeneity, nor what functional form best describes the association between these covariates and treatment effects. A researcher may wish to describe subpopulations that are of interest a priori, and which can be defined by a known combination of covariates. However, as the set of demographic variables increase, analysts that perform post hoc analysis by looking for patterns in the data that were not specified a priori, run into the well-known multiple hypothesis testing problem.

As an example, consumers in different socioeconomic groups and with distinct historical intra-day load profiles, or behavioural characteristics, may respond differently to the introduction of tariffs that charge different prices for electricity at different times of the day. Customers who can (cannot) adapt their consumption profile to TOU tariffs will accrue a benefit (cost). Those who consume electricity at more expensive peak periods, and who are unable to change their consumption patterns, could end up paying significantly more.

In assessing whether demographic variables are informative in terms of the impact of TOU tariffs on load profiles, the Customer-Led Network Revolution project (Sidebotham & Powergrid 2015) noted

> .. a relatively consistent average demand profile across the different demographic groups, with much higher variability *within* groups than *between* them. This high variability is seen both in total consumption and in peak demand.

In addition, the question of which demographic variables are important when considering the impact of energy policies ignores the fact that many of these variables should be considered together, in a multiplicative fashion. One reason for this finding might be that it is the (unknown) combination of income, household size, education, and daily usage patterns that describes a particular vulnerable demographic group.

In this paper we consider the distributional effects on customers following the introduction of Time-of-Use (TOU) pricing schemes where the price per kWh of electricity usage depends on the time of consumption. These pricing schemes are enabled by smart meters, which can regularly (e.g. half-hourly) record consumption. Using machine learning methods, we consider the association between the effect of TOU pricing schemes on household electricity demand and a range of variables that are observable before the introduction of the new pricing schemes. Our chosen method allows the analyst to be agnostic both with respect to which variables are important and the functional form.

We demonstrate the application of a recently developed method, known as a causal tree, and an aggregation of causal tree estimates known as a causal forest (Athey & Imbens 2016, Wager & Athey 2017). These methods search across covariates for good predictors of heterogeneous treatment effects. Causal trees provide an interpretable description of heterogeneity, while causal forests can be used to obtain individual-specific estimates of treatment effects. The Conditional Average Treatment Effect (CATE) estimator, the expected effect of a treatment for individuals in a subpopulation defined by covariates, can be used to obtain estimates of a treatment effect that varies.

Given that policy makers are often interested in the factors underlying a given prediction, it is desirable to gain some insight to which variables in this large set are most often selected. A key challenge follows from that fact that partitions generated by tree-based methods are sensitive to subsampling, while the use of ensemble methods such as causal forests produce more stable, but less interpretable estimates.

To address this problem we utilise variable importance measures to consider which variables are chosen most often by the causal forest algorithm. However, in the estimation of variable importance it is important to account for the impact of the varying information content across continuous versus discrete random variables. In particular, tree based methods can be biased towards continuous variables, given the presence of more potential splitting points. We address this issue by including permutation-based tests for our variable importance results. This is particularly important for this analysis given that many of our demographic variables are either binary or categorical.

In section 2 we first describe the potential outcomes framework and conditional average treatment effects, then describe causal trees and causal forests. In section 3, we discuss issues of interpretability and describe the variable importance measures. In section 4, we introduce the application to electricity smart meter data, and review existing literature. In section 5, we present the results. Section 6 concludes.

2 Methods for Estimation of Heterogeneous Treatment Effects

The estimand is defined using the potential outcomes framework introduced by Neyman (1923) and developed by Rubin (1974). Let X_i be a vector of covariates for individual i. Suppose that there is one treatment group of interest. $Y_i(1)$ ($Y_i(0)$) denotes the potential outcome if individual i is allocated to the treatment (control) group. The causal effect of a treatment on individual i is therefore $Y_i(1) - Y_i(0)$. The fundamental problem of causal inference is that we do not observe the causal effect for any i (Holland 1986).

The estimand that we consider is the Conditional Average Treatment Effect (CATE)

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

Whereas the ATE can be estimated by a difference in means $\bar{y}_t - \bar{y}_c$, where $\bar{y}_t$ ($\bar{y}_c$) is the mean of the outcome variable for the treated (control) group, the CATE can be thought of as a subpopulation average treatment effect.[1] [2] The CATE is identified under unconfoundedness, i.e. $Y_i(1), Y_i(0) \perp T_i|X_i$, and overlap, i.e. $0 < \Pr(T_i = 1|X_i = x) < 1 \; \forall \; x$, where T_i denotes the treatment indicator variable.

The CATE can be operationalised by including interactions between the treatment indicators and the conditioning variable(s) of interest. The inclusion of interaction terms in a linear model is a common technique for exploring the heterogeneity of treatment effects in areas ranging from biomedical science to the social sciences.[3]

It is possible to search for heterogeneity in treatment effects simply by separately estimating CATEs using many possible conditioning variables and repeatedly estimating the standard linear regression model, and conducting tests of multiple hypotheses. However, a clear problem is false discovery and the need to adjust significance levels for multiple hypothesis testing which can limit the power of a test to find heterogeneity.

A number of alternative machine learning methods allow the researcher to explore more complex forms of heterogeneity. Recent methods involving LASSO and treatment effect estimation are described in papers by Imai et al. (2013), Weisberg & Pontes (2015) and Tian et al. (2014). However, Athey & Imbens (2017) note some drawbacks of LASSO methods, particularly the need for sparsity assumptions.

LASSO methods are preferable to tree and forest methods when outcomes or treatment effects are linearly or polynomially related to the covariates. In this study we are interested in allowing for many possibly nonlinear interactions between covariates, which is more easily implementable through forest methods.

2.1 Regression Trees

In this section we provide an overview of the Classification and Regression Tree (CART) method of Breiman et al. (1984). We describe regression trees, and then describe two key adaptations to regression tree methods introduced by Athey & Imbens (2016): honest estimation - the use of separate subsamples for constructing the tree and for obtaining estimates for each leaf, and the adjustment of the splitting criterion for when treatment effects are estimated for each leaf. [4]

Suppose there are p covariates and N observations. The objective is to partition the covariate space $\mathbb{X}$ into M mutually exclusive regions $R_1, ..., R_M$, where the outcome for an individual with covariate vector x in region R_m is estimated as the mean of the outcomes for training observations in leaf R_m. The following algorithm is used to apply binary splits of the data:

Let X_j be a splitting variable and s be a split point. Define $R_1(j, s) = \{X|X_j \leq s\}$ and $R_2(j, s) =$

[1] In instances where we condition on x being in some subset of the covariate space, i.e. $x \in \mathbb{A} \subset \mathbb{X}$, and $\tau_{\mathbb{A}} = E[Y_i(1) - Y_i(0)|x \in \mathbb{A}]$, we also refer to this as the CATE (with suitably re-defined covariates).

[2] Another estimand is the average treatment effect conditional upon observed covariates $\bar{\tau} = \frac{1}{N}\sum_{i=1}^{N} \tau(x_i) = \frac{1}{N}\sum_{i=1}^{N} E[Y_i(1) - Y_i(0)|X_i = x_i]$. Imbens & Rubin (2015) refer to this as the conditional average treatment effect, but we shall use the above definition of the CATE.

[3] A description of the application of linear regression methods for the purpose of estimating treatment effects in randomized experiments can be found in Athey & Imbens (2017).

[4] This section summarizes the description of regression trees provided by Hastie et al. (2009), and the description of honest estimation provided by Athey & Imbens (2016).

$\{X|X_j > s\}$.[5] The algorithm selects the pair (j, s) that solves:

$$\min_{j,s} \left[\sum_{x_i \in R_1(j,s)} (y_i - \bar{y}_1(j,s))^2 + \sum_{x_i \in R_2(j,s)} (y_i - \bar{y}_2(j,s))^2 \right] \tag{2}$$

where $\bar{y}_1(j,s)$ and $\bar{y}_2(j,s)$ are the mean outcomes in $R_1(j,s)$ and $R_2(j,s)$ respectively. When the data has been split into two regions, the same process is applied separately to each region. Then the process is repeated on each of the four resulting regions, and so on.

In machine learning a dataset is often divided into training and testing data, denoted by $\mathcal{S}^{tr}$ and $\mathcal{S}^{te}$ respectively. Model selection, which in the case of a tree is the partition that defines the tree, and estimation are carried out on $\mathcal{S}^{tr}$ with the goal of minimizing expected mean squared error in $\mathcal{S}^{te}$. Often, the selection and estimation of a model also requires a choice of value for some tuning parameter, which can be used to avoid overfitting.

The tuning parameter can be chosen by cross-validation, which involves splitting the training data into training and validation subsamples, respectively $\mathcal{S}^{tr,tr}$ and $\mathcal{S}^{tr,cv}$. The model can be fitted for different parameter values using $\mathcal{S}^{tr,tr}$, with the MSE in $\mathcal{S}^{tr,cv}$ used to evaluate the choice of α. The final chosen α is then used in selection and estimation carried out on all of $\mathcal{S}^{tr}$.

A common approach for limiting the amount of overfitting is to grow a tree T_0, stopping when some minimum node size is reached, and then to "prune" the tree in the following way: A subtree $T \subset T_0$ is any tree that can be obtained by collapsing any number of non-terminal nodes. Let the terminal nodes be indexed by m and let $|T|$ be the number of terminal nodes in T. Let N_m be the number of observations in R_m, and let the splitting criterion be $C_\alpha(T) = \sum_{m=1}^{|T|} \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2 + \alpha|T|$, where $\hat{c}_m = \frac{1}{N_m} \sum_{x_i \in R_m} y_i$. Given α, pruning finds the subtree $T_\alpha \subseteq T_0$ that minimizes $C_\alpha(T)$. The tuning parameter $\alpha \geq 0$ determines the trade-off between tree size and goodness of fit. For the final tree $T_{\hat{\alpha}}$, the value $\hat{\alpha}$ can be chosen such that it minimizes the cross-validated Mean Square Error.

Adaptive and Honest estimation

Let the outcome for individual i be denoted by Y_i and the sample mean for the leaf in which a tree allocates an individual with covariates X_i be denoted by $\hat{\mu}(X_i; \mathcal{S}^{tr}, \Pi(\mathcal{S}^{tr}))$. Π denotes a partition of the covariate space and $\Pi(\mathcal{S}^{tr})$ is a partition created by applying the regression tree algorithm to the training data.

The target for adaptive regression trees is to minimize MSE in test data[6]

$$\mathbf{E}_{\mathcal{S}^{te},\mathcal{S}^{tr}} \left[\text{MSE}_\mu(\mathcal{S}^{te}, \mathcal{S}^{tr}, \Pi) \right] \equiv \mathbf{E}_{\mathcal{S}^{te},\mathcal{S}^{tr}} \left[\frac{1}{N^{te}} \sum_{i \in \mathcal{S}^{te}} \{(Y_i - \hat{\mu}(X_i; \mathcal{S}^{tr}, \pi(\mathcal{S}^{tr})))^2 - Y_i^2\} \right] \tag{3}$$

A standard regression tree is referred to as *adaptive* in order to distinguish it from so-called *honest* regression trees (Athey & Imbens 2016). The adaptive regression tree splitting criterion is given by $\text{MSE}_\mu(\mathcal{S}^{tr}, \mathcal{S}^{tr}, \Pi) + \alpha \times$ *no. of splits*, where the first argument of $\text{MSE}_\mu(.)$ indicates that the error is evaluated in-sample on the training data $\mathcal{S}^{tr}$. The second argument indicates that the leaf means are calculated using the training data $\mathcal{S}^{tr}$. Π is a potential partition of the covariate space.

Standard machine learning methods are biased because they use the same training data for model selection and estimation (see Athey & Imbens (2016)). *Honest* methods avoid this problem by using different information for selecting the model and for estimation. In the context of regression trees, an honest regression tree involves partitioning the training data into separate samples used to construct the tree (i.e. choosing the splits, including cross-validation), and for estimating the within-leaf means. Following the notation of Athey & Imbens (2016), we let $\mathcal{S}^{tr}$ and $\mathcal{S}^{est}$ denote, respectively, the training and estimation subsamples. It should be noted that while this method eliminates the bias and allows for estimates with standard asymptotic properties there is also a potential loss of precision resulting from smaller sample size.

For honest regression trees the target criterion is $\mathbf{E}_{\mathcal{S}^{te},\mathcal{S}^{est},\mathcal{S}^{tr}} \text{MSE}_\mu(\mathcal{S}^{te}, \mathcal{S}^{est}, \Pi(\mathcal{S}^{tr}))$ where $\mathcal{S}^{te}$ indicates that MSE is constructed using test data, and $\mathcal{S}^{est}$ denotes that leaf means will be calculated using

[5] If a splitting variable is categorical with q unordered values, then we can consider all $2^{q-1} - 1$ possible splits of the q values into two groups, or we can use binary variables for each category.

[6] The adjustment Y_i^2 does not affect the ranking of estimators.

independent estimation data. Note that the splits of the tree are chosen in honest estimation without using the data that will be used for estimating leaf means.

A critical difference between adaptive and honest estimators is that the honest splitting criterion takes account of the uncertainty associated with the yet to be constructed leaf-mean estimates. This is accomplished by including an estimate of within-leaf variance, $\frac{1}{N^{est}} \sum_{\ell \in \Pi} S^2_{\mathcal{S}^{tr}}(\ell(x;\Pi))$, where N^{est} is the number of observations in $\mathcal{S}^{est}$.

The estimate of the expected mean square error is

$$\widehat{\text{EMSE}}_{\mu}(\mathcal{S}^{tr}, N^{est}, \Pi) \equiv -\frac{1}{N^{tr}} \sum_{i \in \mathcal{S}^{tr}} \hat{\mu}^2(X_i; \mathcal{S}^{tr}, \Pi) + (\frac{1}{N^{tr}} + \frac{1}{N^{est}}) \sum_{\ell \in \Pi} S^2_{\mathcal{S}^{tr}}(\ell(x;\Pi)) \tag{4}$$

where $S^2_{\mathcal{S}^{tr}}(\ell(x;\Pi))$ is the estimated within-leaf variance. The term $(\frac{1}{N^{tr}} + \frac{1}{N^{est}}) \sum_{\ell \in \Pi} S^2_{\mathcal{S}^{tr}}(\ell(x;\Pi))$ penalizes finer partitions that lead to greater variance in leaf estimates. The adaptive criterion is $-\frac{1}{N^{tr}} \sum_{i \in \mathcal{S}^{tr}} \hat{\mu}^2(X_i; \mathcal{S}^{tr}, \Pi)$.

The splitting criterion is then written as $\widehat{\text{EMSE}}_{\mu}(\mathcal{S}^{tr}, N^{est}, \Pi) + \alpha \times$ *no. of splits*, where the tuning parameter α is chosen using the cross-validation criterion $\widehat{\text{EMSE}}_{\mu}(\mathcal{S}^{tr,cv}, N^{est}, \Pi)$.[7]

2.2 Tree Methods for Estimating Treatment Effects

Causal trees are different to regression trees in that the leaf estimates are CATEs, obtained by a simple difference in means. Whereas regression trees are constructed by recursively splitting the data in order to minimize the mean square error of estimated outcomes, causal tree splits are based on minimizing an estimate of the *infeasible* mean square error of estimated treatment effects. Below we briefly outline a number of approaches that adjust regression tree methods for the treatment effect context.

A straightforward method involves fitting trees separately to treatment and control group individuals (Athey & Imbens 2016, 2015). The estimated treatment effect for any set of covariates is simply the difference in the estimated outcomes for the two trees.[8] However, in this two-tree approach the splits take account of heterogeneity in separate potential outcomes rather than heterogeneity in the *treatment effects*.

Athey & Imbens (2016, 2015) outline an approach that involves using a transformed outcome $Y_i^* = Y_i.(W_i - p)/(p.(1-p))$, where p is the probability of treatment. This Transformed Outcome Tree (TOT) method has the advantage that $\mathbb{E}[Y_i^*|X_i = x] = \tau(x)$ and off-the-shelf regression tree methods can be applied. In general this method is not efficient because the information in the treatment indicator is only used in constructing the transformed outcome. Athey & Imbens (2016) also compare causal trees to methods based on the t-statistic for treatment effect differences (Su et al. 2009), and outcome prediction error (Zeileis et al. 2008).

The preferred method is the causal tree algorithm which utilises the within-leaf difference in sample means for treatment and control groups (Athey & Imbens 2016). This is preferable to fit-based trees or the two-tree method because splitting is based on obtaining more accurate predictions of treatment effects, rather than the separate treatment and control outcomes. The difference-in-means causal tree produces less noisy estimates than the TOT method because it makes more use of the information in the treatment indicator.

Adaptive Causal Trees

The issue of adaptive versus honest estimation applies to both regression trees and causal trees. The adaptive methods use the same data for splitting and constructing leaf estimates: leaf means for regression trees and leaf differences- in-means for causal trees ($\bar{Y}^{\ell}_{treated} - \bar{Y}^{\ell}_{control}$). An adaptive regression tree splits based on in-sample MSE, while an adaptive causal tree splits based on an estimate of the infeasible in-sample MSE.

Let τ_i denote the treatment effect for individual i and $\hat{\tau}(X_i; \mathcal{S}^{est}, \Pi)$ denote the estimate of the average treatment effect for the leaf to which individual i with covariates X_i has been allocated. For causal trees the infeasible test data MSE is $\text{MSE}_{\tau}(\mathcal{S}^{te}, \mathcal{S}^{est}, \Pi) \equiv \frac{1}{N^{te}} \sum_{i \in \mathcal{S}^{te}} \{(\tau_i - \hat{\tau}(X_i; \mathcal{S}^{est}, \Pi))^2 - \tau_i^2\}$. While we

[7] $\widehat{\text{EMSE}}_{\mu}(\mathcal{S}^{tr}, N^{est}, \Pi)$ is an approximately unbiased estimator of $\text{EMSE}_{\mu}(\Pi)$ for a fixed Π. It is not unbiased when repeatedly used to evaluate splits, and therefore it is likely to overstate the goodness of fit for deep trees. Therefore cross-validation still plays a role, albeit a less important role.

[8] Similar methods are used by Beygelzimer & Langford (2009) and Foster et al. (2011).

never know τ_i (the mean-squared error of the treatment effect is thus infeasible), an unbiased estimator of $\text{MSE}_\tau(\mathcal{S}^{te}, \mathcal{S}^{tr}, \Pi)$, can be obtained by recognising the fact that $\hat{\tau}$ is constant within leaves. Expanding $\text{MSE}_\tau(\mathcal{S}^{te}, \mathcal{S}^{tr}, \Pi)$ and then exploiting $\mathbf{E}_{\mathcal{S}^{te}}[\tau_i | i \in \mathcal{S}^{te} : i \in \ell(x, \Pi)] = \mathbf{E}_{\mathcal{S}^{te}}[\hat{\tau}(x; \mathcal{S}^{te}, \Pi)]$, gives

$$\widehat{\text{MSE}}_\tau(\mathcal{S}^{te}, \mathcal{S}^{tr}, \Pi) \equiv -\frac{2}{N^{te}} \sum_{i \in \mathcal{S}^{te}} \hat{\tau}(X_i; \mathcal{S}^{te}, \Pi).\hat{\tau}(X_i; \mathcal{S}^{tr}, \Pi) + \frac{1}{N^{te}} \sum_{i \in \mathcal{S}^{te}} \hat{\tau}^2(X_i; \mathcal{S}^{tr}, \Pi). \tag{5}$$

Given that $\mathcal{S}^{te}$ is unknown when the tree is being constructed,[9] (5) cannot be used as the splitting criterion. If we replace $\hat{\tau}(X_i; \mathcal{S}^{te}, \Pi)$ in (5) with $\hat{\tau}(X_i; \mathcal{S}^{tr}, \Pi)$, this gives an estimator of the infeasible in-sample goodness-of-fit, $\hat{\text{MSE}}_\tau(\mathcal{S}^{tr}, \mathcal{S}^{tr}, \Pi) \equiv -\frac{1}{N^{tr}} \sum_{i \in \mathcal{S}^{tr}} \hat{\tau}^2(X_i; \mathcal{S}^{tr}, \Pi)$, used in the splitting criterion, $\hat{\text{MSE}}_\tau(\mathcal{S}^{tr}, \mathcal{S}^{tr}, \Pi) + \alpha \times \textit{number of splits}$, where α is set by cross-validation. The cross-validation criterion is $\hat{\text{MSE}}_\tau(\mathcal{S}^{tr,cv}, \mathcal{S}^{tr,tr}, \Pi)$.

Adaptive causal trees give biased estimates, and Athey & Imbens (2016) find that unbiased honest causal trees perform better in simulations in terms of MSE and coverage of confidence intervals.

Honest Causal Trees

With the aim of minimizing $\mathbf{E}_{\mathcal{S}^{te}, \mathcal{S}^{est}, \mathcal{S}^{tr}} \text{MSE}_\tau(\mathcal{S}^{te}, \mathcal{S}^{est}, \Pi(\mathcal{S}^{tr}))$, the estimate of the expected MSE used with the honest causal tree splitting criterion is given by

$$\hat{\text{EMSE}}_\tau(\mathcal{S}^{tr}, N^{est}, \Pi) \equiv -\frac{1}{N^{tr}} \sum_{i \in \mathcal{S}^{tr}} \hat{\tau}^2(X_i; \mathcal{S}^{tr}, \Pi) + (\frac{1}{N^{tr}} + \frac{1}{N^{est}}) \sum_{\ell \in \Pi} \left(\frac{S^2_{\mathcal{S}^{tr}_{treat}}(\ell)}{p} + \frac{S^2_{\mathcal{S}^{tr}_{control}}(\ell)}{1-p} \right) \tag{6}$$

where p is the probability of allocation to the treatment group, and $S_{\mathcal{S}^{tr}_{treat}}$ ($S_{\mathcal{S}^{tr}_{control}}$) is the training sample variance for treated (control) observations in leaf ℓ. For determining the penalty parameter, α, by cross-validation, we use $\hat{\text{EMSE}}_\tau(\mathcal{S}^{tr,cv}, N^{est}, \Pi)$.

Some additional parameters must be specified when fitting causal trees. We must specify the minimum number of treatment and control observations required in leaves resulting from a split. If we use honest estimation, then we must decide how much data to use for training and how much to use for estimation.

2.3 Forests

Since individual trees are noisy, forests emerge from averaging over many trees, thereby reducing the variance. The estimates produced by random forests are often more accurate than single tree estimates in terms of MSE. Below we provide a brief description of a random forest.

The prediction of a random forest is the average of many unpruned regression trees. Each tree is produced using a bootstrap sample without replacement. At each split in the tree, the algorithm uses a random subset of the set of all covariates as potential splitting variables. Each tree is fully grown up to a minimum leaf size.

A random forest algorithm (Friedman et al. 2009) proceeds by drawing a bootstrap sample of size N from the training data, and then growing a random tree T_b (b indexing the bootstrap samples). This is accomplished by recursively (until the minimum node size n_{min} is reached) selecting m variables at random from the p variables, and picking the best variable and split point among the m variables. The chosen node is then split into two daughter nodes.

The prediction for an individual with a vector of covariates x is then $\frac{1}{B} \sum_{b=1}^{B} T_b(x)$, where $T_b(x)$ is the estimate produced by tree b. The trees are not independent since two bootstrap samples can have some common observations, and therefore the correlation between trees limits the benefits of averaging. However, this correlation is reduced through the random selection of the input variables.

Similar aggregations over causal trees, known as causal forests, can improve the accuracy of treatment effect estimates. Wager & Athey (2017) outline the properties of causal forests and show that, under certain assumptions, the predictions from causal forests are asymptotically normal and centred on the true treatment effect for each individual. Recent applications of causal forests can be found in papers by Davis & Heller (2017*a*,*b*) and Bertrand et al. (2017). The forests in these papers use an honest splitting rule for the construction of the causal trees.

[9] We note that this issue applies to both standard regression and causal trees.

3 Interpretation of Causal Forest Estimates

A general issue which applies to standard regression trees and random forests is the trade-off between interpretability and stability. A single causal tree splits the data into relatively few leaves. The results are easy to interpret given that a simple tree diagram allows the researcher to quickly identify the subgroup to which any household belongs by following a set of decision rules. Strobl (2008) notes that single trees can be unstable with small changes in the training data resulting in a very different model (tree). However, although stable forests generate 'better' predictive performance, the interpretability of a single tree is lost when we move to an ensemble method, such as a causal forest.

Across the many trees within a forest, it is not immediately clear what covariates most strongly influence the final estimates, and how different covariates interact. This follows given that the set of splitting variables can be used with different splitting points, and in different combinations. Given that applied econometricians are often interested in the factors underlying a given prediction, it is therefore desirable to gain some insight to which variables in this large set are most often selected by the causal forest output.

To do this we utilise variable importance measures to consider which variables are chosen most often by the causal forest algorithm. However, in the estimation of variable importance it is important to account for the impact of the varying information content across continuous versus discrete random variables. In particular, tree based methods can be biased towards continuous variables, given the presence of more potential splitting points. We address this issue by including permutation-based tests for our variable importance results.

3.1 Variable Importance

A standard measure of variable importance first proposed by Breiman et al. (1984) uses, for variable ℓ, the sum of improvements in squared error brought about by splits where the splitting rule uses variable ℓ. For decision tree T, with $J-1$ internal nodes, the importance of variable ℓ in tree T is given by

$$\mathcal{I}_\ell^2(T) = \sum_{t=1}^{J-1} \hat{i}_t^2 I(v(t) = \ell), \tag{7}$$

where $\hat{i}_t^2$ is the estimated improvement in squared error at node t, $I()$ is an indicator function, and $v(t)$ is the variable chosen at node t that gives the maximal estimated improvement in squared error at that node (Hastie et al. 2009).[10] It is standard practice to assign a value of 100 to the most important variable and scale the measures for the other variables accordingly.

This measure is applied to random forests (or any additive tree expansions) by averaging over M trees, giving $\mathcal{I}_\ell^2 = \frac{1}{M}\sum_{m=1}^{M} \mathcal{I}_\ell^2(T_m)$. Hastie et al. (2009) note that "due to the stabilizing effect of averaging, this measure turns out to be more reliable than its counterpart for a single tree". As noted by Breiman et al. (1984) and Strobl (2008), this measure is biased towards variables with a higher number of categories and continuous variables because these variables have more potential splitting points. Variables can be incorrectly split on because one of many possible split points is spuriously found to reduce the most error in the training data.

Variable Importance for Causal Forests

In the application of variable importance for causal forests, we note that the "ground truth" treatment effect for any individual is unobservable. However, it is possible to implement a method similar to the standard squared error loss variable importance measure described above. For honest causal forests, we can use the improvement in the honest splitting criterion. The measure also takes surrogate splits into account, i.e. it takes account of splits that are not carried out on the variable of interest, but which are similar to potential splits on the variable of interest. When a variable is a surrogate for a splitting variable, this approach adds to the variable's tree importance the concordance of that surrogate with the

[10] This measure is often also adjusted to take account of improvements in fit for nodes at which the variable of interest is a good surrogate for the splitting variable (Breiman et al. 1984). This addresses the potential problem of the masking of the importance of variables that are not chosen for a split, but are highly correlated to the splitting variable.

splitting variable multiplied by the improvement from the split. This reduces masking of variables that are not used for a split, but that are correlated with the splitting variable.[11]

The aforementioned bias of variable importance measures towards continuous variables and variables with many categories can be avoided by making use of discretized variables with equal numbers of categories.[12] However, discretization of variables can also lead to a loss of useful information, and reduce the accuracy of our estimates. Another standard variable importance measure is based upon a count of the proportion of splits on the variable of interest up to a depth of 4, with a depth-specific weighting.[13] [14]

$$imp(x_j) = \frac{\sum_{k=1}^{4} \left[\frac{\sum_{all\ trees} number\ depth\ k\ splits\ on\ x_j}{\sum_{all\ trees} total\ number\ depth\ k\ splits} \right] k^{-2}}{\sum_{k=1}^{4} k^{-2}} \tag{8}$$

3.2 Permutation Test for Causal Forest Variable Importance

If the splits in trees spuriously occur more often on continuous variables and variables with more categories, then this should also occur when the dependent variable is permuted. In this instance, the p-value should be unaffected unless the extent of the over-selection of variables for splitting is also dependent on the true importance of the variables. We investigate this issue in further detail in Appendix A, which contains a simple simulation study of this permutation based variable importance test. The simulations suggest that the p-values are potentially unaffected by the bias of variable splitting towards variables with more possible splitting points.

Following the method of Altmann et al. (2010) for random forests,[15] and Bleich et al. (2014) for BART, we compute p-values for the default variable importances provided by the `grf` package. This involves permuting the dependent variable 1000 times and obtaining variable importances for all variables from 1000 causal forests fitted separately using the 1000 permutations as dependent variables. The variable importances are also obtained from a causal forest using the original, unpermuted dependent variable. Then, following the "local" test described by Bleich et al. (2014), we obtain a p-value for each variable by finding the proportion of the 1000 causal forests for which the variable had a greater variable importance measure than that obtained from the causal forest with the unpermuted dependent variable.

4 Heterogeneity of Household Electricity Demand Response

TOU tariffs are becoming more implementable through the use of smart metering technology. Understanding heterogeneity in household responses to TOU pricing is of interest to both regulators and retailers. The subsequent increase in the availability of large amounts of past electricity consumption data allows for more household specific targeting of electricity pricing and other demand stimuli. Furthermore, in a world where energy suppliers rely increasingly on renewables which are intermittent in nature, measures to reduce peak demand are required as part of the need to balance supply and demand.

The British energy regulator, Ofgem (2013), is interested in the impact of new pricing schemes upon vulnerable and low income customers. Faruqui et al. (2010) postulate that two potentially offsetting forces influence how we expect low-income customers to be impacted differently by new electricity pricing schemes. First, lower income customers can have a greater proportion of their demand in off-peak hours, and therefore can benefit from TOU pricing without adjusting their daily demand profile. Second, we might not expect these customers to shift and reduce load as much as other customers because they

[11] The measure is provided for individual causal trees in the `R` package `causalTree`. This follows the approach used in the regression tree `R` package `rpart`.

[12] This approach can be implemented through an option provided by Athey et al. (2016) in the `R` package `causalTree`. The authors include an option to determine splits by separately ordering treated and untreated individuals according to a potential splitting variable, then putting observations into numbered buckets, with a minimum number of buckets and a maximum bucket size.

[13] This is the default measure for the command `causal_forest` in the `R` package `grf`.

[14] In order to obtain variable importances for categorical variables, which currently must be entered into the `causal_forest` command as a set of binary variables for each level of the categorical variable, we take the sum of the variable importances of the binary variables. The parameters we set for the `causal_forest` command are: 15000 trees, bootstrap samples of half the data, one third of covariates randomly drawn as potential splitting variables for each split, and target minimum node size of 5.

[15] Altmann et al. (2010) show that p-values based on permutation of the dependent variable can address the issues of bias towards variables with more categories, and masking of the importance of groups of highly correlated variables.

have lower usage levels in general and less discretionary usage. The authors confirm these hypotheses using US data, and find that low income customers change their electricity usage less than higher income customers.

Counter to some of this evidence, studies by Lower Carbon London (Schofield et al. 2014) and Frontier Economics and Sustainability First (DECC 2012) have noted the generally low associations between demographic variables and demand response, and in particular, the lack of evidence pertaining to differing responses of low-income and vulnerable customers. One possible reason for this is that the nature of the heterogeneity is additive. Individuals most affected by energy policies might be identified through the interaction of a number of variables. For example, the Centre for Sustainable Energy produced a report (Preston et al. 2013) which used interactions of variables to define the groups of households predicted to face the largest increase in household bills as a result of changes in energy policy.

In this study we examine the importance of variables constructed from historic load profiles. Relatively few studies have conditioned upon past usage data when estimating treatment effects of electricity pricing schemes. Some recent examples include a study using US data by Harding & Lamarche (2016), who split the sample into low, medium, and high usage customers. The results suggest that high usage customers decrease peak usage to a greater extent, which is somewhat expected since these customers have more reducible usage. However, surprisingly low-income customers appear to increase consumption in off-peak time periods. The authors speculate that this substantial load-shifting by low-income customers is the result of moral licencing and note that this indicates the difficulty in anticipating the impact of new pricing schemes for some customer segments. A number of recent studies have used past electricity usage data for the estimation of household-specific treatment effects. Bollinger & Hartmann (2015) condition upon the empirical distribution of past electricity usage and consider how a utility can gain from targeting based upon ITE estimates. Balandat (2016) estimates ITEs by comparing predictions of electricity usage under control group allocation to realised usage under treatment allocation during the trial period.

Data

The dataset used in this project is from the Electricity Smart Metering Customer Behavioural Trial conducted by the Irish Commission for Energy Regulation (CER 2011). The CER note that this is "one of the largest and most statistically robust smart metering behavioural trials conducted internationally to date" (CER 2011). The dataset consists of half hourly residential electricity demand observations for 4225 households over 536 days. The benchmark period began on 14th July 2009 and ended on 31st December 2009. Households were then randomly allocated to either a control group or various TOU Pricing Schemes and Demand Side Management stimuli from 1st January 2010 to 31st December 2010.

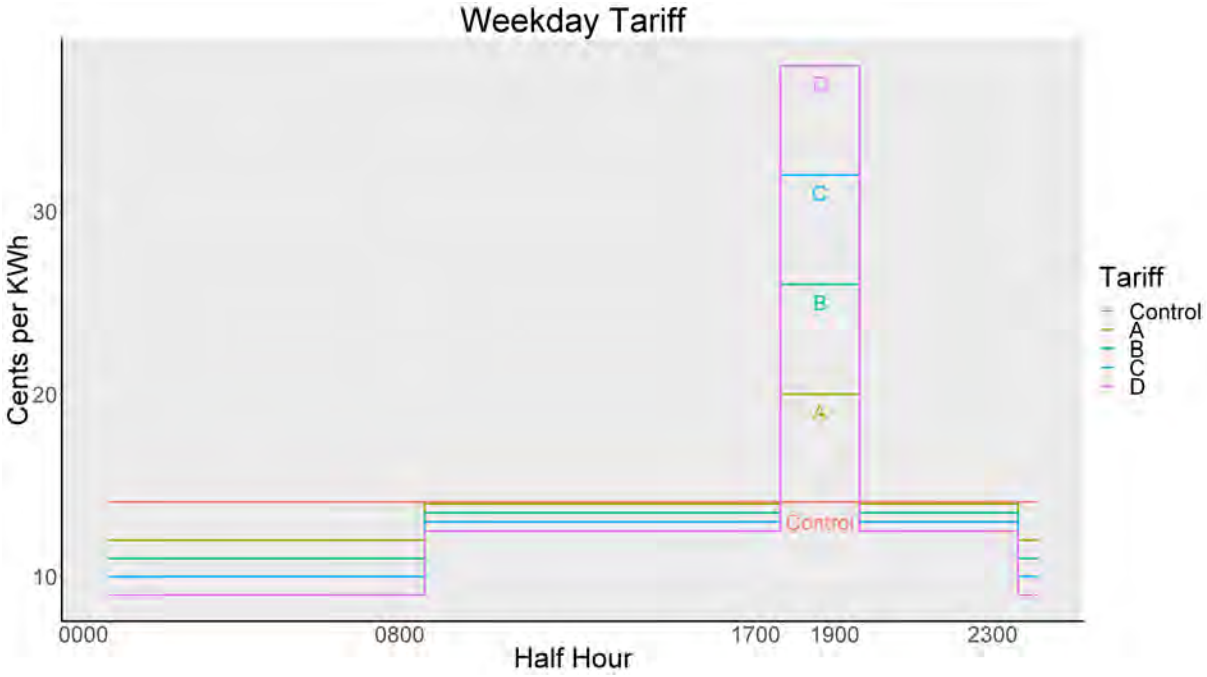
All households were charged the normal Electric Ireland tariff of 14.1 cents per kWh during the benchmark period. During the trial period the control group remained on the tariff of 14.1 cents per kWh while the test group were allocated to tariffs A, B, C, or D.[16] The tariffs A to D were structured as shown in Table 1, and are graphed in Figure 1a.

Table 1: TOU Tariff details

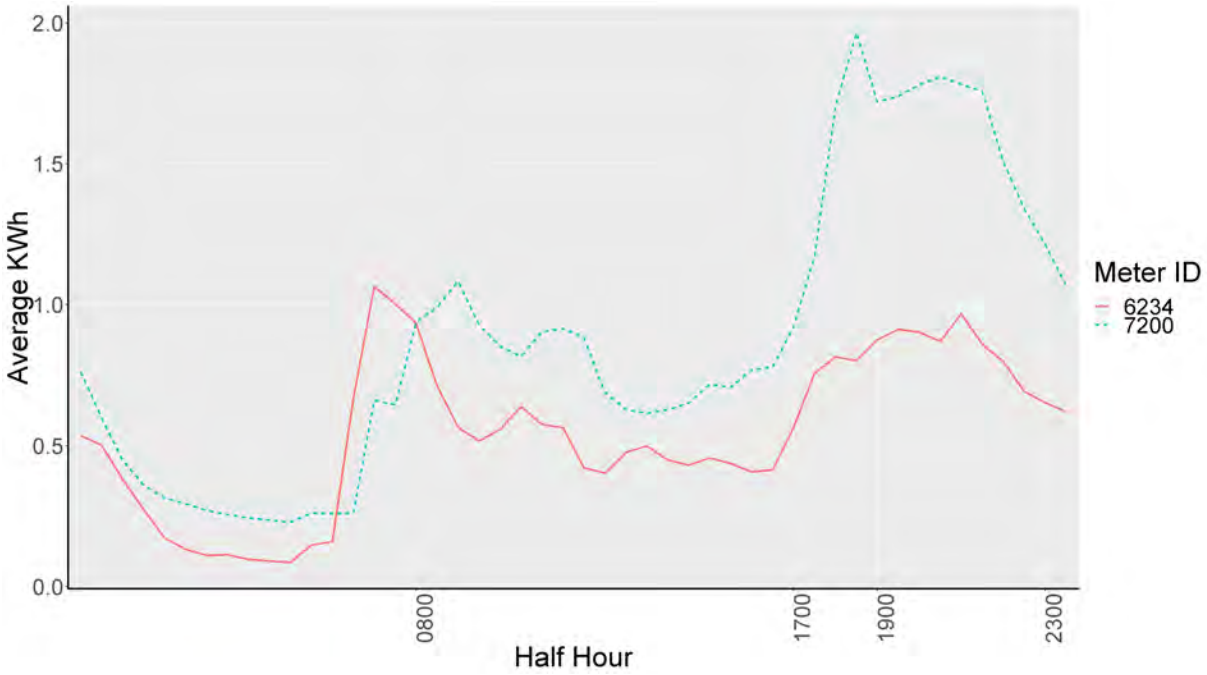
TOU Tariffs (cents per kWh)	**Night** 23.00-08.00	**Day** 08.00-17.00 every day 19.00-23.00 every day 17.00-19.00 weekends and holidays	**Peak** 17.00-19.00 Mon-Fri Excluding holidays
Tariff A	12.00	14.00	20.00
Tariff B	11.00	13.50	26.00
Tariff C	10.00	13.00	32.00
Tariff D	9.00	12.50	38.00

Households in the test group were also allocated to one of the following Demand Side Management (DSM) stimuli: Bi-monthly detailed Bill; Monthly detailed bill; Bi-monthly detailed bill and In-Home Display (IHD); Bi-monthly detailed bill and Overall Load Reduction (OLR) incentive.

[16] There was also a Weekend tariff group, which we exclude from this study.



(a) Trial period TOU tariffs



(b) Pre-trial average half-hourly demand for two households

Figure 1: Prices and examples of demand profiles

The identification of ATEs depends upon unconfoundedness and overlap. The CER took a number of steps to ensure that the samples for treatment groups were representative and did not exhibit notable biases. A stratified random sampling framework was used with phased recruitment. Non-respondents and attriters were surveyed and adjustments were made accordingly. Those who opted in were compared to the national profile. The full dataset contains 4225 households, with 768 households in the control group and 233 households facing the combination of tariff C and IHD stimulus, which will be the treatment group of interest in this paper.

Figure 1b gives an example of average half hourly usage on weekdays before the trial period for households with similar survey responses. The two households both have four people in a 3 bedroom semi-detached house, in which the chief earner is an employee and lower middle class with 3rd level education. Both households also typically have one person at home during the day, own their home, have timed oil heating, and have a similar stock of appliances. This figure shows that even households that are similar across multiple characteristics do not necessarily have the same patterns of demand use. Therefore survey variables are limited in describing demand heterogeneity.[17]

5 Results

The outcome variable is average half-hourly peak time electricity consumption during the trial period (measured in kWh), excluding weekends. We restrict attention to Tariff C in combination with the In-Home Display (IHD). The IHD stimulus is of greater interest than the other information stimuli, and tariff C has a high ratio of peak to off-peak prices and more observations than any other tariff combined with the IHD.[18]

Below we present two estimates of single causal trees as an example of the instability of single tree estimates and small sample size. Causal forest Individual Treatment Effect (ITE) estimates are then described in terms of their association with pre-trial variables. Finally, variable importance measures are presented in order to consider which variables are the strongest determinants of the structure of the trees in the forest.

The standard ATE estimates for the tariff C with IHD range from -0.073 to -0.092 kWh for an average peak half hour, depending on the set of controls.[19] Mean half-hourly peak consumption for the control group during the trial period (one full year) was 0.799 kWh, while mean peak consumption for all households during the pre-trial period (half a year) was 0.828 kWh. Therefore these treatment effects are of the order of 10% of peak consumption.

5.1 Causal Trees

Figures 2 and 3 show estimated honest causal trees. The set of potential splitting variables is given in Table 2. The minimum number of treatment and control observations required for a leaf split is set to ten. Half of the data is used for creating the splits in the tree, and half is used for honest estimation. The only difference in estimation of the two trees is the seed for random number generation, which determines the subsampling of the data into splitting and estimation data, and determines subsamples used for cross-validation. The diagrams contain 90% confidence intervals.

It can be immediately observed from these trees that the partition of the data generated by the causal tree algorithm is sensitive to the input data. This can be viewed as partly a sample size issue. Sample size, in combination with sample splitting for honest estimation, also has implications for statistical significance. There were 500 observations used for splitting, and 501 observations for estimation of treatment effects. The causal tree output contains few subgroups with significantly non-zero treatment effects at the 5% level. In contrast, CATE estimates obtained from a low-variance method, such as a linear model interacting treatment with different levels of education and including control variables, can result in multiple groups with significant effects.

[17] In this paper we make use of pre-trial survey data, but we cautiously avoid using post-trial survey information. Prest (2017) applies an adjusted causal tree method to this data, but the estimates are potentially biased by conditioning on post-trial survey information. Our methods also differ from those of Prest (2017) in that we make use of a forest, which should lead to estimates that are more stable with respect to training data.

[18] 343 households were allocated to Tariff C with the IHD, whereas only 126 households were allocated to tariff D with the IHD.

[19] These results are obtained by linear regression of average peak usage on the treatment indication.

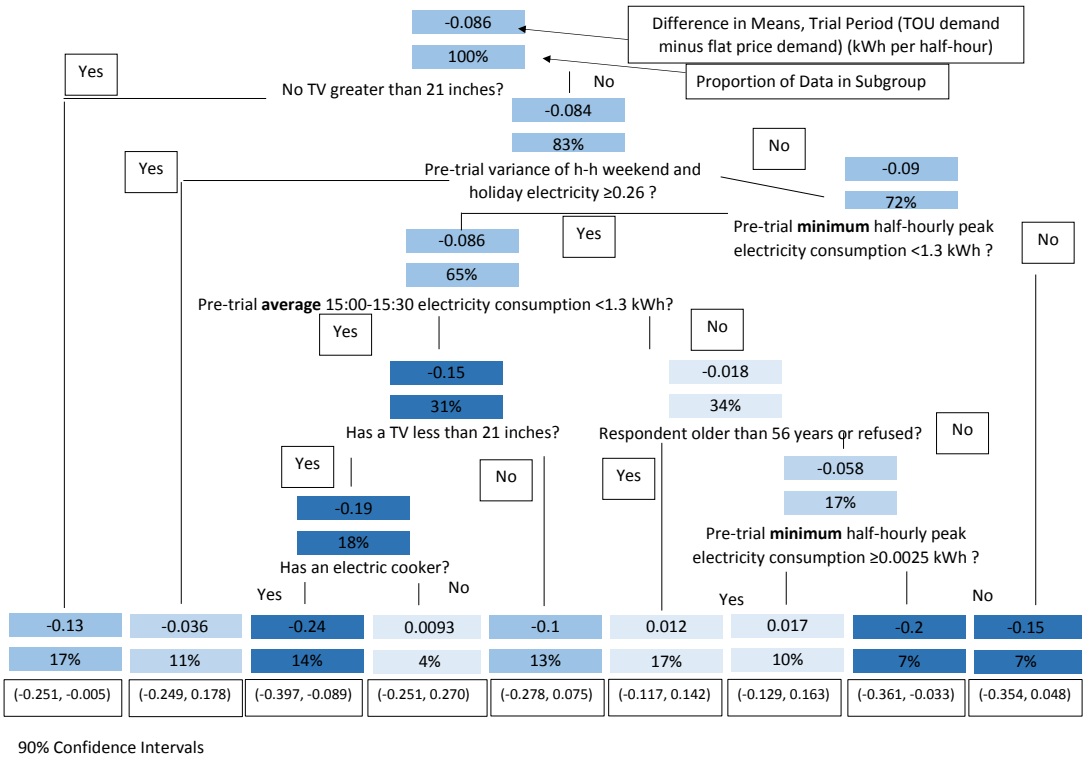


Figure 2: Single Tree Example 1

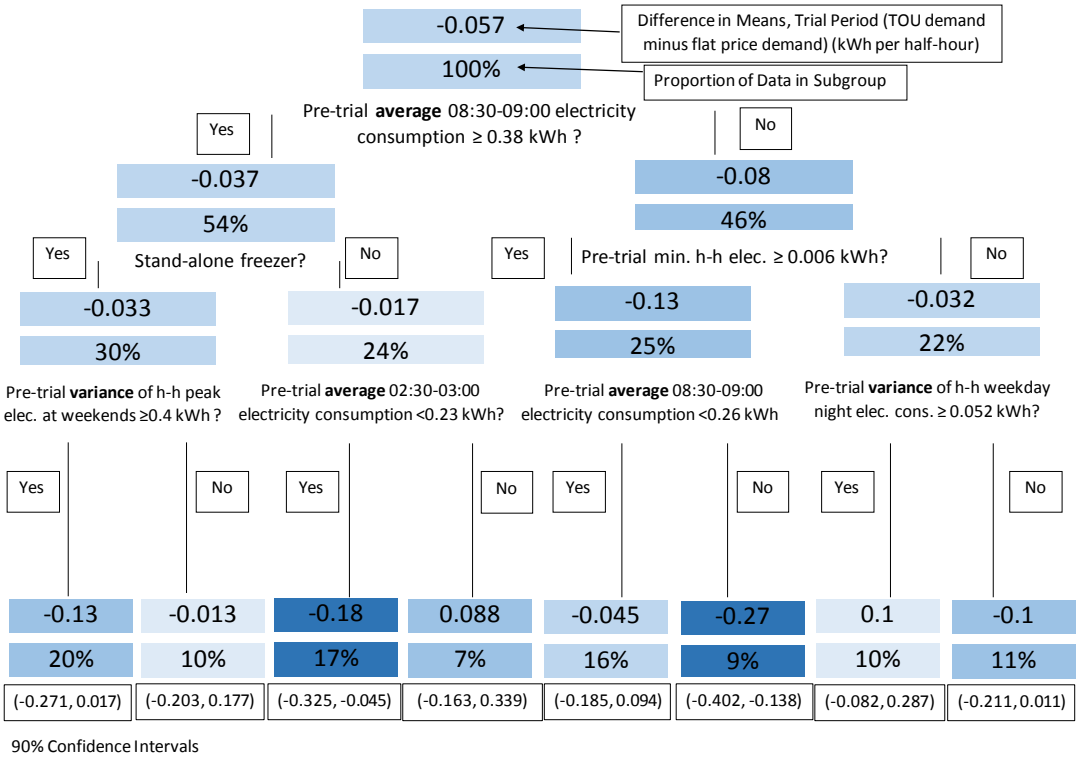


Figure 3: Single Tree Example 2 - Different seed

The above instability can be addressed by the use of a causal forest. The instability of the output (i.e. sensitivity to the random separation of the data into splitting and estimation subsamples) is less of a problem when aggregation of predictions occurs over a large number of honest causal trees.

5.2 Causal Forest

We fitted a causal forest to the dataset containing a set of control households and households allocated to tariff C and the IHD stimulus (1001 households). Each individual honest tree is fitted using a bootstrap sample consisting of half of the data, with half of this sample used for splitting and half used for estimation.[20] The number of individual trees fitted is 15000.[21] For each tree in the forest, a random subsample of one third of the set of covariates are used as potential splitting variables.[22] The minimum number of treatment and control observations required for a leaf split is set to five.

In order to determine if our estimates of demand response give a reasonable characterisation of heterogeneity, in Table 3 we present averages of past consumption variables for each quartile of ITE estimates. Table 4 contains binary survey variables and gives the percentage of observations in each quartile of ITE estimates. For example, we can observe that for the first quartile of treatment effects, i.e. the quartile of most responsive households, 40% of households have a respondent with third level education.

For the vast majority of covariates we observe associations across quantiles of individual effects that we would expect a priori. The most responsive households (i.e. Quartile 1) generally use more electricity, are more educated, younger, higher social class, and have more appliances. This particular result is in agreement with the observation made by Di Cosmo et al. (2014), using the same data, that more educated households are generally more responsive.[23]

Tables 5 and 6 present an overview of the association between demographic covariates and the quartiles of ITE estimates. These tables give the percentages of all households in different combinations between quartiles and age categories. For example, 9.8% of households are in the fourth quartile of treatment effects and have a respondent aged over 65. Demographic groups that are more likely to contain vulnerable customers (CSE 2012), namely lower class and retired households, together with households for which the respondent was over 65 years old, contain a greater proportion of less responsive households (see Table 5). While this may be largely due to the fact that these groups have less reducible peak usage, this difference in demand response for vulnerable and non-vulnerable groups could be relevant to regulation of potential consumer targeting.

The patterns of heterogeneity observed in both Tables 3 and 4 are largely maintained when the forest is fitted using only electricity consumption data.[24] To demonstrate this, Figure 4a presents a density plot comparing the distributions of the ITE estimates obtained by fitting causal forests with different sets of potential conditioning variables. One forest was fitted using both survey and usage variables, one forest was fitted using only usage variables, and one forest was fitted using only survey variables. This suggests that electricity consumption data contains information related to survey data information that can characterise heterogeneous groups of demand response. This issue may be relevant to firms or policymakers who wish to understand which information to collect in order to predict demand response.

The results suggest that the usage variables exert a greater influence on the causal forest estimates. Furthermore, the density plot suggests potential bimodality in the distribution of individual effects which is not noticeable from the estimates produced by using survey variables alone. However, while it is most plausible that past usage variables are more informative than survey variables, we must also consider the possibility that these results are driven by the bias of variable selection towards continuous variables, which have more potential splitting points. Figure 4b gives a similar comparison of density plots of ITE estimates, but where estimates are produced from causal forests with tree splits determined by the

[20] Bertrand et al. (2017) also use these sizes of bootstrap samples and training and estimation subsamples. Wager & Athey (2017) divide bootstrap samples in half for honest estimation.

[21] This is somewhat arbitrary, and between the values of 10000 and 25000 used by Bertrand et al. (2017) and Davis & Heller (2017*b*).

[22] Random Forests and Causal Forests should randomly subsample a set of potential conditioning variables at each split within each tree, but the `causalForest` command in the `R` package `causalTree` currently only supports sampling splitting variables for each tree, and the results are likely to be similar. The choice of one third of the total number of covariates is commonly used for random forests.

[23] Our focus on peak demand response is also justified by the observation by Di Cosmo & O'Hora (2017) that households "reduced consumption rather than shifting consumption from peak".

[24] The results for causal forests fitted using only survey variables or only usage variables are not included in this paper, but are available from the authors on request.

Table 2: Potential splitting variables for Causal Trees and Causal Forest

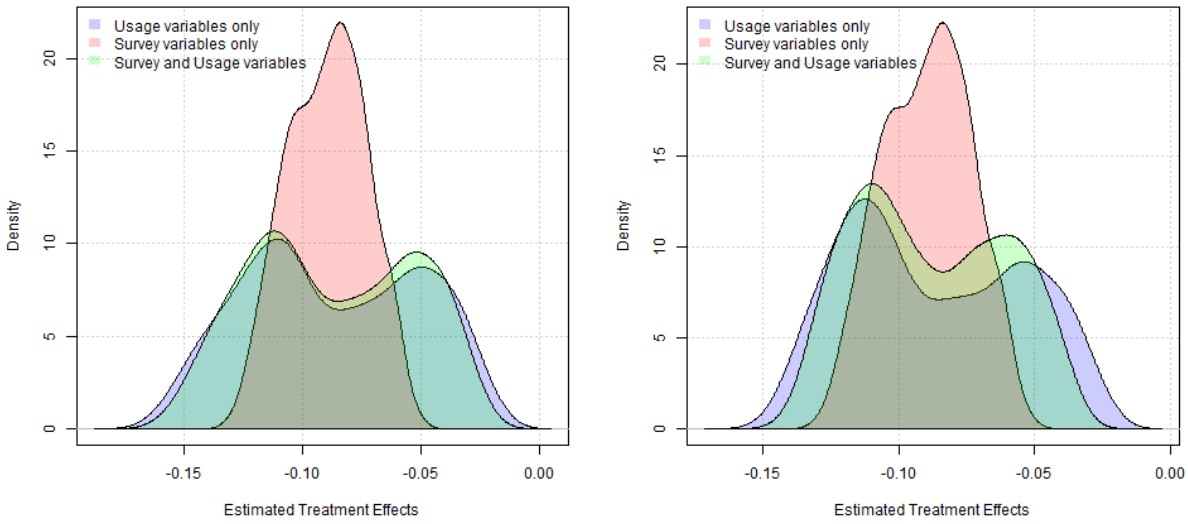
Name of variable	
Survey variables (categorical)	
Age of respondent	Sex of respondent
Class of chief income earner	Regular internet use
Employment status of chief income earner	Other reg. internet users
Number of bedrooms	Education of chief earner
Type of home	Electric central heating
Alone or other occupants	Electric plugin heating
Own or rent the home	Central water heating
Number of electric cookers - number	Immersion water heating
Internet access	Instant water heating
Approximate age of home	Number of washing machines
Lack money for heating	Number of tumble dryers
Number of dishwashers	Number of instant electric showers
No. showers elec. pumped from hot tank	Type of cooker
Number of plug-in convector heaters	Number of freezers
Number of water pumps or electric wells	Number of immersion water heaters
Number of small TVs	Number of big TVs
Number of desktop PCs	Number of laptop PCs
Number of games consoles	Has an energy rating
Proportion of energy saving lightbulbs	Prop. double glazed windows
Lagging jacket	Attic insulation
External walls insulated	
Electricity usage variables (continuous)	
Mean usage	Min. usage
Variance of usage	Max. usage
Mean peak usage	Mean nonpeak usage
Variance of peak usage	Variance of nonpeak usage
Mean night usage	Mean daytime usage
Variance of night usage	Variance of daytime usage
Mean usage - weekdays	Mean peak usage - weekdays
Variance of usage - weekdays	Var. peak usage - weekdays
Mean night usage - weekdays	Mean daytime usage - weekdays
Variance of night usage - weekdays	Var. daytime usage - weekdays
Mean daily maximum usage	Mean usage - weekends
Mean daily minimum usage	Variance of usage - weekends
Mean of half-hour coefficients of variation	Mean usage - each month (July-Dec)
Avg. night usage/ avg. daily usage	Var. of usage - each month (July-Dec)
Avg. lunchtime usage/ Avg. daily usage	Mean usage - each half-hour
Mean night usage - weekends	Mean daytime usage - weekends
Variance of night usage - weekends	Var. daytime usage - weekends

Table 3: Pre-trial electricity consumption variable averages for quartiles of causal forest estimates of household Treatment Effect

	Quartile of Estimated TE on Peak Usage			
Variable	Q1	Q2	Q3	Q4
Predicted TE (kWh)	-0.13	-0.10	-0.07	-0.04
Avg. pre-trial half-hourly usage (kWh)	0.72	0.64	0.40	0.23
Avg. pre-trial peak half-hourly usage (kWh)	1.35	1.02	0.62	0.35
Var. of pre-trial half-hourly usage (kWh)	0.70	0.51	0.27	0.11
Var. pre-trial peak half-hourly usage (kWh)	1.23	0.79	0.42	0.19
Max half-hour elec. con. (kWh)	7.42	6.58	5.34	3.87
Min half-hour elec. cons. (kWh)	0.03	0.04	0.02	0.01
Mean daily max (kWh)	3.43	2.90	2.15	1.30
Mean daily min (kWh)	0.12	0.14	0.07	0.04

Table 4: Binary survey variable averages for quartiles of causal forest estimates of household Treatment Effect

	Quartile of Estimated TE on Peak Usage			
Variable	Q1	Q2	Q3	Q4
Male	52%	54%	53%	48%
Internet access	86%	80%	57%	43%
Elec. central heating	3.2%	4.4%	5.2%	4.8%
Water immersion	61%	65%	50%	44%
Water centrally heated	13%	17%	14%	11%
Went without heat from lack of money	4.4%	3.6%	2.8%	3.6%
Lagging jacket on hot water	85%	83%	86%	77%
Higher Education	40%	39%	34%	28%
Employee	56%	49%	39%	33%
Apartment	0%	0.8%	2%	5.2%
Instantaneous water heater	0.8%	0.4%	1.6%	2%
Plug-in electric heater	2.8%	4%	4.8%	2.8%



(a) Standard splitting

(b) Bucket splitting method

Figure 4: Density plots of causal forest household estimates fitted using different sets of variables

Table 5: Percentages of households in combinations of survey categories and treatment effect quartiles

Variable	*Quartile of Estimated TE on Peak Usage*			
Age	Q1	Q2	Q3	Q4
18-25	0.1%	0%	0.1%	0.1%
26 - 35	2.3%	2.5%	2.8%	2.2%
36 - 45	6.0%	5.3%	3.5%	3.9%
46 - 55	8.3%	6.1%	4.6%	4.4%
56 - 65	5.8%	5.1%	4.6%	4.4%
65+	2.4%	5.7%	9.4%	9.8%
Refused	0.2%	0.3%	0%	0.2%
Class	Q1	Q2	Q3	Q4
Upper middle and middle	4.4%	4.3%	1.8%	1.5%
Lower middle	7.4%	6.4%	6.6%	5.3%
Skilled working	4.7%	4.9%	4.8%	3.2%
Working and non-working	8.0%	8.6%	10.6%	14.1%
Farmers	0.5%	0.5%	0.7%	0.8%
Refused	0.1%	0.3%	0.5%	0.1%
Employment	Q1	Q2	Q3	Q4
Employee	14.0%	12.3%	9.8%	8.3%
Self-emp (with employees)	1.8%	2.4%	0.7%	0.3%
Self-emp (with no employees)	1.8%	1.2%	1.5%	0.8%
Unemployed (seeking work)	1.6%	0.2%	1.1%	1.8%
Unemployed (not seeking work)	1.0%	0.8%	0.6%	1.1%
Retired	4.4%	7.9%	11.2%	12.4%
Carer	0.5%	0.2%	0.1%	0.3%
Education	Q1	Q2	Q3	Q4
No formal education	0.4%	0.2%	0.4%	0.4%
Primary	2.2%	2.9%	3.0%	5.4%
Secondary - junior cert	3.7%	4.3%	3.9%	4.5%
Secondary - leaving cert	7.6%	6.5%	8.2%	6.2%
Third level	10.1%	9.7%	8.4%	7.0%
Refused	1.1%	1.4%	1.1%	1.5%
Residents	Q1	Q2	Q3	Q4
Lives Alone	0.5%	2.0%	6.3%	13.5%
All people over 15	13.0%	15.1%	14.8%	9.5%
Both adults and children	11.6%	7.9%	3.9%	2.0%
Number of bedrooms	Q1	Q2	Q3	Q4
1	0%	0.3%	0.2%	1.2%
2	0.4%	1.3%	2.2%	6.0%
3	8.3%	8.8%	13.2%	12.2%
4	11.5%	11.4%	7.7%	4.3%
5+	4.9%	3.1%	1.6%	1.3%
Refused	0%	0.1%	0.1%	0%
Own or rent	Q1	Q2	Q3	Q4
Rent (private landlord)	0.3%	0.2%	0.5%	0.8%
Rent (local authority)	1.0%	0.8%	0.9%	2.1%
Own Outright	12.6%	12.9%	16.0%	15.3%
Own with mortgage	11.2%	11.0%	7.5%	6.6%
Other	0%	0.1%	0.1%	0.2%

Table 6: Percentages of households in combination of survey categories and treatment effect quartiles - Appliance variables

Variable	*Quartile of Estimated TE on Peak Usage*			
Number of washing machines	Q1	Q2	Q3	Q4
None	0.1%	0.3%	0.1%	1.3%
One	24.5%	24.4%	24.8%	23.6%
Two	0.5%	0.3%	0.1%	0.1%
Number of tumble dryers	Q1	Q2	Q3	Q4
None	2.3%	5.9%	9.3%	14.8%
One	22.6%	19.0%	15.7%	10.2%
Two	0.2%	0.1%	0%	0%
Number of Dishwashers	Q1	Q2	Q3	Q4
None	3.5%	5.1%	10.7%	16.1%
One	21.5%	19.9%	14.3%	8.9%
Two	0.1%	0%	0%	0%
Number of instant elec. showers	Q1	Q2	Q3	Q4
None	6.2%	6.5%	7.6%	11.0%
One	16.5%	16.8%	16.6%	13.3%
Two	1.9%	1.4%	0.8%	0.7%
More than Two	0.5%	0.3%	0%	0%
Number of Electric Cookers	Q1	Q2	Q3	Q4
None	2.9%	4.7%	5.6%	9.5%
One	22.1%	20.3%	19.3%	15.5%
Two	0.1%	0%	0.1%	0%
Immersion	Q1	Q2	Q3	Q4
None	4.4%	5.1%	6.3%	8.5%
One	20.5%	19.9%	18.7%	16.4%
Two	0.2%	0%	0%	0.1%
Number of large TVs	Q1	Q2	Q3	Q4
None	3.0%	3.3%	4.9%	7.8%
One	11.3%	11.1%	13.4%	13.3%
Two	8.3%	6.6%	6.3%	3.3%
Three	2.0%	2.8%	0.4%	0.5%
More than three	0.5%	1.2%	0%	0.1%
Number of laptop PCs	Q1	Q2	Q3	Q4
None	8.9%	9.5%	13.8%	16.2%
One	11.8%	11.7%	10.0%	8.2%
Two	3.4%	2.4%	0.8%	0.5%
Three	0.8%	1.0%	0.3%	0%
More than three	0.2%	0.4%	0.1%	0.1%
Approx. prop. saving lightbulbs	Q1	Q2	Q3	Q4
None	4.3%	5.2%	5.5%	7.5%
A quarter	7.1%	6.0%	6.1%	5.7%
A half	4.5%	4.5%	4.5%	4.1%
Three quarters	5.1%	4.9%	4.1%	3.4%
All	4.1%	4.4%	4.8%	4.3%

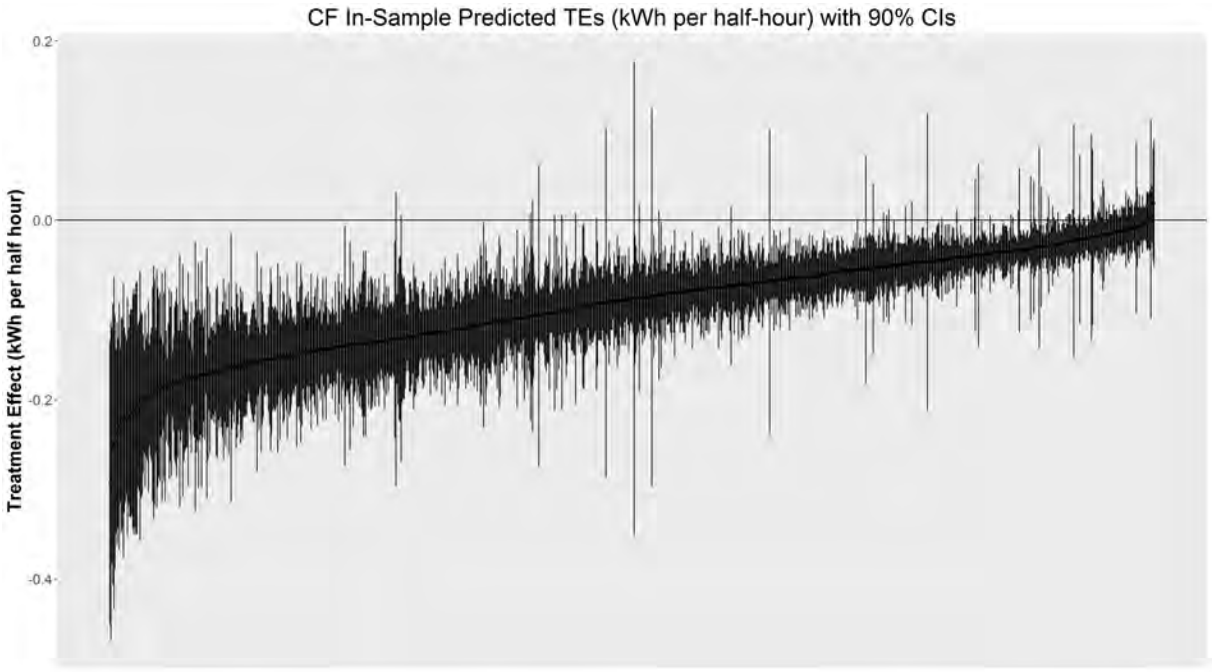


Figure 5: 90% Confidence Intervals for ITEs ordered by size of ITE

bucket splitting method described in Section 3. The overall shape of the density plots indicates that the importance of our historic usage variables is robust to any bias that may originate from the higher number of potential splitting points, relative to the demographic variables.

Figure 5 shows ITEs with confidence intervals ordered by size of estimated effect.[25] None of the individual estimates are significantly positive. This accords with economic intuition.

5.3 Variable Importance

In this section we present the results for variable importance utilising the methods outlined in Section 3. The first method utilises the average improvement in the causal tree splitting criterion from splits on the variable of interest,[26] and the second method is a depth-weighted average of the number of splits on the variable of interest.[27] For the second method we also carry out a permutation-based test, as outlined in section 3.

Columns (1) and (2) of Table 7 give the variable names and values for the variable importance measure which utilises the improvement in the causal tree splitting criterion. The variables are ordered by importance, with larger values indicating greater importance. The results indicate that the trees most often split on electricity usage, and specifically variables that indicate the level and variance of weekday electricity consumption (i.e. mean peak usage - weekdays, var. night usage). The most important survey variables are employment status and a variable for the number of electric pumped showers (*employment*, *number of hot tank elec. showers*).

Columns (3) and (4) of Table 7 give names and values for the variable importances based upon the number of splits obtained from a causal forest with the unpermuted dependent variable. These results are similar to the variable importance measures in column 2, but more strongly favour the continuous electricity usage variables.

As noted in Section 3, given the bias of variable importance measures in favour of variables with more splitting points (Strobl 2008), we implement an alternative *permutation* test of variable importance which

[25] These confidence intervals are produced by the `causal_forest` command of the R package `grf`. See Wager & Athey (2017) for a description of how these intervals are constructed. Each level of a categorical survey variable is represented by a separate binary potential splitting variable because the package currently does not support finding optimal splits of multiple categories.

[26] This is default measure in the `R` package `causalTree`.

[27] This is the default measure in the `R` package `grf`.

(1) causalForest variable importance	(2)	(3) grf variable importance	(4)	(5) p-value	(1) causalForest variable importance	(2)	(3) grf variable importance	(4)	(5) p-value
attic insulated	0.04	*water instantly heated*	0	0.9	mean 13:00-13:30 usage	40.68	*number of freezers*	10.02	0.09
mean 01:00-01:30 usage	0.08	*number of washing machines*	0.17	0.95	mean 07:00-07:30 usage	40.77	mean h-h coef. of variation	10.51	1
mean 00:30-01:00 usage	0.1	*unheated, lack of money*	0.22	0.78	var. September peak usage	40.78	mean daytime usage	10.83	0.19
prop. elec. saving lightbulbs	0.14	*electric plugin heating*	0.25	0.27	*number of bedrooms*	40.9	variance night usage	11.03	0.98
mean 07:30-08:00 usage	0.16	*electric central heating*	0.34	0.95	mean 15:00-15:30 usage	41.1	mean 10:30-11:00 usage	11.33	0.94
mean usage - weekdays	0.86	*prop. double glazed windows*	0.42	1	*own or rent home*	41.21	mean 22:00-22:30 usage	11.4	0.76
mean 00:00-00:30 usage	1.3	*number of electric cookers*	0.52	1	mean 21:00-21:30 usage	41.43	mean 13:00-13:30 usage	11.82	0.86
variance daytime usage	1.37	*number of tumble dryers*	0.59	1	variance peak usage	42.06	var. night usage - weekdays	11.86	0.99
external walls insulated	1.51	*number of dishwashers*	0.73	1	mean 10:30-11:00 usage	42.13	mean 23:00-23:30 usage	12.09	0.93
mean 08:00-08:30 usage	1.8	*number of immersion heaters*	0.81	1	*number of small TVs*	42.9	mean 14:30-15:00 usage	12.11	0.8
mean 05:00-05:30 usage	1.89	*sex of respondent*	1.08	1	*type of home*	43.36	var. night usage - weekends	12.17	0.98
variance nonpeak usage	2.03	*type of cooker*	1.08	1	*electric central heating*	44.66	mean 21:30-22:00 usage	12.26	0.63
mean h-h coef. of variation	2.12	*attic insulated*	1.12	1	*education*	44.82	*number of laptop PCs*	12.56	0.19
lagging jacking	2.31	*own or rent home*	1.21	1	mean peak usage	45	mean 22:30-23:00 usage	12.7	0.81
mean 04:00-04:30 usage	2.33	*no. of elec. convector heaters*	1.22	1	mean 14:30-15:00 usage	45.12	mean 06:30-07:00 usage	12.72	0.97
mean 05:30-06:00 usage	2.53	*regular internet user*	1.24	1	mean night usage	45.36	mean daytime usage - weekends	12.78	0.1
mean daytime usage	2.56	*water pumped from elec. well*	1.4	1	*number of dishwashers*	45.38	mean 00:00-00:30 usage	13.66	0.89
mean 02:00-02:30 usage	2.81	*water immersion*	1.41	0.99	mean 12:30-13:00 usage	45.4	mean daytime usage - weekdays	14.16	0.12
no. of elec. convector heaters	3.19	*number of instant elec. showers*	1.47	1	*other internet users*	45.44	variance nonpeak usage	14.23	0.19
water pumped from elec. well	3.31	*other internet users*	1.48	0.61	mean daily max. usage	45.54	var. nonpeak usage - weekdays	15	0.26
mean 06:00-06:30 usage	3.4	*external walls insulated*	1.49	1	var. December peak usage	45.84	mean daily min. usage	15.84	0.9
number of desktop PCs	3.42	*number of hot tank elec. showers*	1.63	1	mean 10:00-10:30 usage	46.8	mean 10:00-10:30 usage	15.89	0.64
mean 03:30-04:00 usage	3.75	*water centrally heated*	2.12	0.98	*electric plugin heating*	46.85	mean 23:30-00:00 usage	15.9	0.78
min. half-hourly usage	3.86	*lagging jacking*	2.16	0.74	mean 12:00-12:30 usage	47.19	mean 07:30-08:00 usage	16.37	0.98
number of freezers	4.02	*age of home*	2.39	1	mean 21:30-22:00 usage	47.5	min. half-hourly usage	16.51	0.88
number of instant elec. showers	4.39	*has an energy rating*	2.85	0.6	mean 11:00-11:30 usage	48.6	variance daytime usage	16.58	0.14
variance of usage	4.91	*number of small TVs*	3.01	1	*lives alone*	48.65	mean lunchtime / mean day usage	16.61	1
number of big TVs	4.96	*number of games consoles*	3.29	0.85	mean 11:30-12:00 usage	50.24	mean 18:00-18:30 usage	16.82	0.34
number of games consoles	5.1	*lives alone*	3.39	0.82	mean 22:00-22:30 usage	50.95	var. daytime usage - weekdays	17.6	0.18
prop. double glazed windows	5.64	mean 02:30-03:00 usage	4.03	1	*unheated, lack of money*	51.56	mean 21:00-21:30 usage	17.61	0.26
max. half-hourly usage	5.73	*type of home*	4.06	1	var. October peak usage	51.7	mean 09:00-09:30 usage	18	0.69
mean 08:30-09:00 usage	6.29	*age of respondent*	4.25	1	*internet access*	52.08	variance of usage	18.14	0.05
var. usage - weekdays	6.52	*education*	4.26	1	*water centrally heated*	52.25	var. usage - weekdays	18.29	0.06
mean 03:00-03:30 usage	7.72	mean 12:00-12:30 usage	4.28	1	mean 16:30-17:00 usage	52.6	max. half-hourly usage	18.53	0.87
has an energy rating	8.97	*number of bedrooms*	4.52	0.96	mean 22:30-23:00 usage	52.7	mean 19:00-19:30 usage	18.67	0.21
mean 01:30-02:00 usage	9.69	*prop. elec. saving lightbulbs*	4.56	1	*type of cooker*	53.21	mean 19:30-20:00 usage	19.41	0.15
mean nonpeak usage	9.69	*internet access*	4.94	0.1	*water instantly heated*	53.31	mean 16:00-16:30 usage	19.46	0.44
mean of usage	10.08	mean 03:30-04:00 usage	4.96	1	*regular internet user*	53.37	mean 20:00-20:30 usage	20.3	0.08
number of laptop PCs	10.44	mean 06:00-06:30 usage	5.3	1	*water immersion*	54.64	mean 15:00-15:30 usage	21.12	0.28
mean 09:00-09:30 usage	12.29	mean 03:00-03:30 usage	5.4	1	mean 23:00-23:30 usage	54.82	var. usage - weekends	21.89	0.08
mean 02:30-03:00 usage	15.9	mean 00:30-01:00 usage	5.7	1	var. July peak usage	55.12	mean November peak usage	22.02	0.18
var. night usage - weekends	23.13	mean 05:30-06:00 usage	6.01	1	*number of electric cookers*	55.44	mean 18:30-19:00 usage	22.3	0.1
mean 04:30-05:00 usage	25.78	mean 04:30-05:00 usage	6.03	1	mean 23:30-00:00 usage	57.26	mean 08:00-08:30 usage	22.37	0.69
mean 16:00-16:30 usage	26.07	mean 01:30-02:00 usage	6.29	1	*number of immersion heaters*	57.53	mean 09:30-10:00 usage	23.8	0.37
mean 17:00-17:30 usage	27.02	mean 11:00-11:30 usage	6.46	1	mean night / mean day usage	59.33	var. daytime usage - weekends	23.94	0.06
mean daily min. usage	28.03	mean 04:00-04:30 usage	6.54	1	mean December peak usage	59.96	mean 16:30-17:00 usage	24.27	0.36
mean 17:30-18:00 usage	28.56	mean 05:00-05:30 usage	6.73	1	*social class*	61.34	var. November peak usage	25.8	0.3
mean 18:30-19:00 usage	28.87	*number of desktop PCs*	7.16	0.12	mean October peak usage	62.19	mean 15:30-16:00 usage	27.1	0.17
mean 18:00-18:30 usage	29.28	mean night usage - weekends	7.24	0.97	mean night usage - weekends	62.44	mean daily max. usage	27.36	0.04
variance night usage	29.51	*social class*	7.51	0.7	var. daytime usage - weekdays	62.5	mean 08:30-09:00 usage	30.15	0.33
mean July peak usage	29.74	*number of big TVs*	7.76	0.53	var. nonpeak usage - weekdays	64.71	mean peak usage - weekdays	33.35	0.03
mean 06:30-07:00 usage	30.99	mean 01:00-01:30 usage	7.91	1	*number of tumble dryers*	71	mean peak usage	34.52	0.01
mean 15:30-16:00 usage	31.99	*employment*	7.93	0.57	*age of respondent*	71.25	mean 20:30-21:00 usage	36.62	0.01
mean September peak usage	32.25	mean 11:30-12:00 usage	8.1	0.99	mean lunchtime / mean day usage	71.33	variance peak usage	40.62	0.01
mean 19:00-19:30 usage	32.69	mean 02:00-02:30 usage	8.1	0.98	*sex of respondent*	71.68	var. peak usage - weekdays	40.73	0.05
mean November peak usage	32.81	mean 12:30-13:00 usage	8.15	1	mean nonpeak usage - weekdays	74.69	var. December peak usage	40.75	0.1
var. August peak usage	33.2	mean night usage	8.2	0.88	*number of hot tank elec. showers*	75.34	mean September peak usage	47.41	0.02
number of washing machines	33.87	mean night usage - weekdays	8.88	0.91	mean usage - weekends	78.1	mean 17:00-17:30 usage	52.63	0.01
mean 20:00-20:30 usage	34.29	mean of usage	8.99	0.18	*employment*	78.23	mean December peak usage	53.13	0
mean 13:30-14:00 usage	35.56	mean night / mean day usage	9.09	1	var. daytime usage - weekends	80.36	mean July peak usage	53.33	0.03
mean 19:30-20:00 usage	35.69	mean nonpeak usage - weekdays	9.14	0.29	mean daytime usage - weekends	82.81	mean 17:30-18:00 usage	53.36	0
var. November peak usage	35.92	mean nonpeak usage	9.38	0.22	mean night usage - weekdays	85.08	mean August peak usage	54.18	0.01
mean 09:30-10:00 usage	36.4	mean 13:30-14:00 usage	9.41	0.95	var. peak usage - weekdays	87.83	var. July peak usage	55.36	0.1
mean 20:30-21:00 usage	36.68	mean 14:00-14:30 usage	9.41	0.92	var. usage - weekends	91.21	var. September peak usage	56.17	0.04
mean 14:00-14:30 usage	36.97	mean usage - weekdays	9.57	0.19	mean daytime usage - weekdays	94.64	mean October peak usage	64.96	0
mean August peak usage	39.59	mean 07:00-07:30 usage	9.91	1	var. night usage - weekdays	95.61	var. August peak usage	71.73	0
age of home	40.15	mean usage - weekends	10.02	0.23	mean peak usage - weekdays	100	var. October peak usage	100	0

Survey variables are in italics.

Table 7: Variable Importance results

is able to address this issue (Altmann et al. 2010). Column (5) shows the p-values for the permutation tests on the `grf` variable importances in column (4).[28] The p-values confirm the pattern of results observed in column (4).

6 Conclusion

In this article we have examined heterogeneity of demand response following the introduction of time-of-use electricity pricing. Tree based methods have a number of advantages relative to other methods that can be applied to this task. The issue of choosing between interpretable but unstable single trees and stable but less interpretable forests with stronger predictive performance is a known issue in the application of standard classification and regression trees.

Variable importance measures, adjusted for differences in information content across past usage and

[28]The variable importances in column 2 of Table 7 are obtained from improvements in the splitting criterion using `causalForest` from the `causalTree` package. However, for computing p-values, we instead use the default variable importance measure provided for `causal_forest` in the `grf` package to increase computational speed.

demographic variables, suggest that the causal forest algorithm favours the use of certain functions of past electricity consumption rather than survey information to describe heterogeneity. Tables 4 to 6 reveal notable patterns of heterogeneity across *unimportant* survey variables. For example, the causal forest results suggest that younger, more educated households that consume more electricity exhibit greater demand response to new pricing schemes. In this respect, although survey variables can be less informative than detailed electricity consumption information in terms of selection in the causal forest algorithm, they can also be correlated with *important* past consumption information.

In Tables 3 to 6 we presented summary statistics on the relationship between a number of individual covariates and the estimated treatment effect of time-of-use electricity prices. One problem with these statistics is a risk of finding spuriously significant results due to multiple hypothesis testing and post-hoc searching across these covariates. Although this problem can be mitigated by restricting attention to a few covariates specified a priori and utilising the methods proposed by Chernozhukov et al. (2017) to conduct valid inference on features of the CATE function, part of the motivation for methods such as causal trees is that the methods can find unknown drivers of heterogeneity. Therefore there is a challenge in combining, on the one hand, avoidance of problems of post-hoc multiple hypothesis testing when attempting to obtain valid inference on descriptions of heterogeneous ITEs, and on the other hand making use of the ability of machine learning methods to discover unknown drivers of heterogeneity from large sets of covariates.[29] Ideally, future research would describe an approach that can discover the key drivers of heterogeneity, and then still provide valid inference on features of the CATE related to these variables.

[29] While variable importance can directly make use of the search for drivers of heterogeneity carried out in binary splitting, other approaches include applying further regression or classification methods on the ITE estimates, for example in articles by Foster et al. (2011), Powers et al. (2017) and Hahn et al. (2017).

References

Altmann, A., Toloşi, L., Sander, O. & Lengauer, T. (2010), 'Permutation importance: a corrected feature importance measure', *Bioinformatics* **26**(10), 1340–1347.

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

Athey, S., Imbens, G., Kong, Y. & Ramachandra, V. (2016), 'An introduction to recursive partitioning for heterogeneous causal effects estimation using causaltree package'.

Athey, S. & Imbens, G. W. (2015), 'Machine learning methods for estimating heterogeneous causal effects', *stat* **1050**(5).

Athey, S. & Imbens, G. W. (2017), 'The econometrics of randomized experiments', *Handbook of Economic Field Experiments* .

Balandat, M. (2016), 'New tools for econometric analysis of high-frequency time series data-application to demand-side management in electricity markets'.

Bertrand, M., Crépon, B., Marguerie, A. & Premand, P. (2017), 'Contemporaneous and post-program impacts of a public works program: Evidence from côte d'ivoire'.

Beygelzimer, A. & Langford, J. (2009), The offset tree for learning with partial labels, *in* 'Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, pp. 129–138.

Bleich, J., Kapelner, A., George, E. I. & Jensen, S. T. (2014), 'Variable selection for bart: An application to gene regulation', *The Annals of Applied Statistics* pp. 1750–1781.

Bollinger, B. & Hartmann, W. R. (2015), Welfare effects of home automation technology with dynamic pricing, Technical report.

Breiman, L., Friedman, J., Stone, C. J. & Olshen, R. A. (1984), *Classification and regression trees*, CRC press.

CER (2011), Electricity smart metering customer behaviour trials (cbt) findings report, Technical report, Commision for Energy Regulation.

Chernozhukov, V., Demirer, M., Duflo, E. & Fernandez-Val, I. (2017), 'Generic machine learning inference on heterogenous treatment effects in randomized experiments', *arXiv preprint arXiv:1712.04802* .

CSE (2012), "beyond average consumption" - development of a framework for assessing impact of policy proposals on different consumer groups, Final report to ofgem, Centre for Sustainable Energy.

Davis, J. & Heller, S. B. (2017*a*), Rethinking the benefits of youth employment programs: The heterogeneous effects of summer jobs, Technical report, National Bureau of Economic Research.

Davis, J. M. & Heller, S. B. (2017*b*), 'Using causal forests to predict treatment heterogeneity: An application to summer jobs', *American Economic Review* **107**(5), 546–550.

DECC (2012), Demand side response in the domestic sector - a literature review of major trials, Technical report, Frontier Economics and Sustainability First, London.

Di Cosmo, V., Lyons, S. & Nolan, A. (2014), 'Estimating the impact of time-of-use pricing on irish electricity demand', *The Energy Journal* **35**(3).

Di Cosmo, V. & O'Hora, D. (2017), 'Nudging electricity consumption using tou pricing and feedback: evidence from irish households', *Journal of Economic Psychology* .

Faruqui, A., Sergici, S. & Palmer, J. (2010), 'The impact of dynamic pricing on low income customers', *Institute for Electric Efficiency Whitepaper* .

Foster, J. C., Taylor, J. M. & Ruberg, S. J. (2011), 'Subgroup identification from randomized clinical trial data', *Statistics in medicine* **30**(24), 2867–2880.

Friedman, J., Hastie, T. & Tibshirani, R. (2009), *The elements of statistical learning*, Vol. 1, Springer series in statistics New York.

Hahn, P. R., Murray, J. S. & Carvalho, C. M. (2017), 'Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects'.

Harding, M. & Lamarche, C. (2016), 'Empowering consumers through data and smart technology: Experimental evidence on the consequences of time-of-use electricity pricing policies', *Journal of Policy Analysis and Management* **35**(4), 906–931.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), Overview of supervised learning, *in* 'The elements of statistical learning', Springer, pp. 9–41.

Holland, P. W. (1986), 'Statistics and causal inference', *Journal of the American statistical Association* **81**(396), 945–960.

Imai, K., Ratkovic, M. et al. (2013), 'Estimating treatment effect heterogeneity in randomized program evaluation', *The Annals of Applied Statistics* **7**(1), 443–470.

Imbens, G. W. & Rubin, D. B. (2015), *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press.

Neyman, J. (1923), 'Sur les applications de la theorie des probabilites aux experiences agricoles: essai des principes (masters thesis); justification of applications of the calculus of probabilities to the solutions of certain questions in agricultural experimentation. excerpts english translation (reprinted)', *Stat Sci* **5**, 463–472.

Ofgem (2013), 'Consumer vulnerability strategy'.
URL: *https://www.ofgem.gov.uk/ofgem-publications/75550/consumer-vulnerability-strategy.pdf*

Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T. & Tibshirani, R. (2017), 'Some methods for heterogeneous treatment effect estimation in high-dimensions', *arXiv preprint arXiv:1707.00102* .

Prest, B. C. (2017), Peaking interest: How awareness drives the effectiveness of time-of-use electricity pricing, *in* 'Riding the Energy Cycles, 35th USAEE/IAEE North American Conference, Nov 12-15, 2017', International Association for Energy Economics.

Preston, I., White, V. & Sturtevant, E. (2013), 'The hardest hit: Going beyond the mean', *a Centre for Sustainable Energy report for Consumer Futures. Available here: http://www. consumerfutures. org. uk/files/2013/05/The-hardest-hit. pdf* .

Rubin, D. B. (1974), 'Estimating causal effects of treatments in randomized and nonrandomized studies.', *Journal of educational Psychology* **66**(5), 688.

Schofield, J., Carmichael, R., amd M. Woolf, S. T., Bilton, M. & Strbac, G. (2014), Residential consumer responsiveness to time-varying pricing, Report a3 for the low carbon london lcnf project, Imperial College London.

Sidebotham, L. & Powergrid, N. (2015), 'Customer-led network revolution project closedown report', *Customer Led Network Revolution. Newcastle upon Tyne* .

Strobl, C. (2008), *Statistical issues in machine learning: Towards reliable split selection and variable importance measures*, Cuvillier Verlag.

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M. & Li, B. (2009), 'Subgroup analysis via recursive partitioning', *Journal of Machine Learning Research* **10**(Feb), 141–158.

Tian, L., Alizadeh, A. A., Gentles, A. J. & Tibshirani, R. (2014), 'A simple method for estimating interactions between a treatment and a large number of covariates', *Journal of the American Statistical Association* **109**(508), 1517–1532.

Wager, S. & Athey, S. (2017), 'Estimation and inference of heterogeneous treatment effects using random forests', *Journal of the American Statistical Association* (just-accepted).

Weisberg, H. I. & Pontes, V. P. (2015), 'Post hoc subgroups in clinical trials: Anathema or analytics?', *Clinical trials* **12**(4), 357–364.

Zeileis, A., Hothorn, T. & Hornik, K. (2008), 'Model-based recursive partitioning', *Journal of Computational and Graphical Statistics* **17**(2), 492–514.

A Simulation Study - Variable Importance Permutation Test

We present a simulation study investigating the extent to which p-values for a permutation-based variable importance test are influenced by the bias of the variable importance measure towards continuous variables and categorical variables with more categories. This study is designed in a similar way to that used by Strobl (2008) for investigating the bias of random forest variable importance measures.

First, we generate the following covariates and treatment indicator: $X_1 \sim N(0,1)$, $X_2 \sim Cat(2)$, $X_3 \sim Cat(4)$, $X_4 \sim Cat(10)$, $X_5 \sim Cat(20)$, $treatment \sim Cat(2)$, where $Cat(k)$ denotes a categorical distribution with k categories of equal probability. Then we consider simulations of the outcome under the following three model designs:

For design 1, none of the covariates affect the outcome, and the outcome is normally distributed: $Y \sim N(0,1)$ For design 2 and 3, the dependent variable is defined in a similar way to a simulation study carried out by Athey & Imbens (2016):

$$Y = \eta(X) + \frac{1}{2}(2 \times treatment - 1) \times \kappa(X) + \epsilon$$

where $\epsilon \sim N(0,1)$. For design 2 the functions are $\eta(X) = 0$ and $\kappa(X) = X_2$, and for design 3 the functions are $\eta(X) = \frac{1}{2}X_1 + X_2$ and $\kappa(X) = X_2$.

We simulate these designs 100 times, with 500 observations per simulation, and for each simulation we permute the dependent variable 100 times and obtain p-values, and then present boxplots of the p-values for each variable.[30] The boxplots of variable importances obtained using the unpermuted dependent variable are shown in Figure 6. The boxplots for the p-values are shown in Figure 7. The boxes give the lower quartile, median, and upper quartiles across repeated simulations. The whiskers give the most extreme data points that are no more than 1.5 times the interquartile range from the box. The circles denote outliers.

Note that the results in Figures 6 and 7 should be interpreted differently. The variable importances in Figure 6 are not used in a test of significance, but rather in a comparison of importance across variables. In contrast, Figure 7 is clearer and correctly indicates that the binary variable is significant in designs 2 and 3. This is an argument in favour of the permutation test.

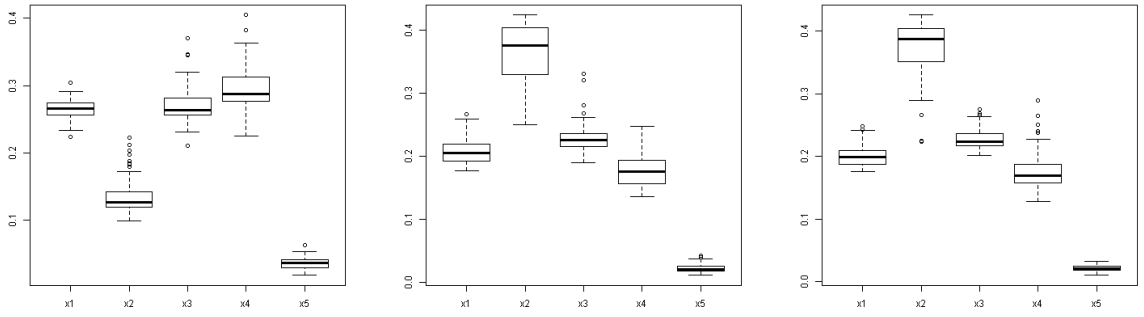
Although for design 1 none of the variables affects the outcome, in Figure 6a X_1 has greater variable importance than X_2, because of the aforementioned bias towards continuous variables.

For categorical variables X_3, X_4, and X_5, all with more categories than X_2, there are two factors influencing the bias of the variable importance measure. As the number of categories increases, there are more potential splits on the variable of interest, because there is a binary variable for each category. This explains why X_3 has greater variable importance than X_2 in Figure 6a. On the other hand, considering the case of a variable with a large number of categories, X_5, there will be relatively few observations allocated to any one category, and therefore a split on one of the X_5 categories is unlikely to lead to a large improvement in the splitting criterion. Therefore the variable importance measures for X_5 are small.

The p-values in Figure 7 appear to be unaffected by these biases. In Figure 7a, none of the variables tend to have significant p-values, reflecting the fact that none of the variables has any influence on the outcome.

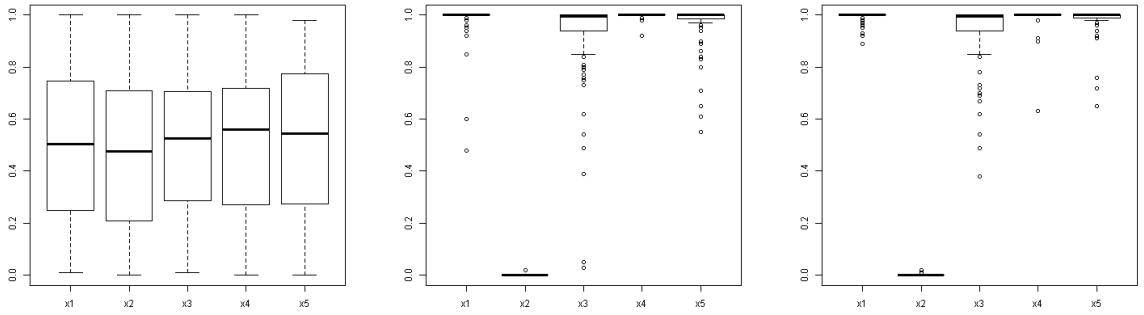
In Figures 7b and 7c, X_2 is correctly identified as the important variable. Although Figures 6b and 6c also indicate that X_2 is the most important variable, there are also misleading differences in the importances of the other variables. However, in Figures 7b and 7c, the variables X_1, X_3, X_4, and X_5 tend to have similar, insignificant p-values.

[30] The parameters for the causal forest are: Number of trees = 5000, bootstrap sample fraction = 0.5, number of potential splitting variables random selected at each split = number of variables divided by 3 and rounded down, minimum node size = 5.



(a) Design 1 var. imp. (b) Design 2 var. imp. (c) Design 3 var. imp.

Figure 6: Boxplots of simulation study variable importances, 100 permutations, 100 iterations



(a) Design 1 p-values (b) Design 2 p-values (c) Design 3 p-values

Figure 7: Boxplots of simulation study p-values, 100 permutations, 100 iterations