Identifying air conditioners using monthly household electricity consumption

Evan D. Sherwin

Friday 28th June, 2019

Background. Residential electricity consumption data contain a wealth of information about both household-level, and electric power system-level characteristics. A growing literature attempts to use such data to understand changes in household electricity consumption behavior. However, such data often lack important household characteristics, such as the presence of air conditioning (AC), a major driver of summer electricity consumption.
Aim. We aim to identify the presence of AC in households on the basis of monthly electricity consumption data, which is available from the billing records of most US electric utilities.
Data. We use aggregate and appliance-level measurements of electricity consumption, in kWh, from 482 households in and around Austin, Texas, of which 449 have AC. We couple this with local temperature data. Electricity consumption readings are available at one-minute resolution, which we aggregate to monthly electricity consumption to emulate widely available monthly data.
Methods. We identify the presence of AC using three methods. First, we fit a line to electricity consumption and monthly average temperature for warmer months of the year and classify households based on the slope of that line. Second, we apply random forest classification, using mean-normalized electricity consumption in each of the twelve months of one year as features. Third, we apply k-means clustering to the above dataset and create a classifier from these clusters.
Results. The linear and random forest classifiers perform comparably out-of-sample, with 97% positive predictive value (PPV) in each case, and 67% and 71% negative predictive value (NPV) respectively. The clustering analysis finds that households without AC tend to have relatively constant electricity consumption throughout the year. The resulting clusters yield a classifier with out-of-sample PPV of 98%, and NPV of 40%.
Conclusions. These results demonstrate that monthly household electricity consumption is a strong predictor of the presence of AC, at least among these households in Austin, Texas. Improvements in NPV, the ability to predict the absence of air conditioners, are likely necessary before these methods can be incorporated into econometric research of electricity consumption behavior. Flexible models, such as the random forest have strong potential. However, a simple linear model has similar performance and likely greater external validity. This line of inquiry shows promise, but larger datasets are likely needed before these methods can be used by econometricians and electric utilities, e.g. to analyze AC adoption patterns.
Keywords: Air conditioner, HVAC, smartmeter, electricity, residential, classification

1 Introduction

Households account for 37% of all US electricity consumption EIA (2016), corresponding to roughly 21% of US greenhouse gas emissions EPA (2016). Understanding determinants of electricity consumption is a key component of any attempt to promote efficiency. Residential electricity consumption data, such as those available to electric utilities, contain a wealth of information about both household-level and electric power system-level characteristics. A growing literature in the machine learning community attempts to learn these characteristics, often deriving an approximate breakdown of household electricity consumption by end use from smartmeter data, with hourly resolution or better Wytock and Kolter (2014); Giri and Berges (2015); Baker et al. (2016). These techniques, known as non-intrusive load monitoring (NILM), are useful both as a tool for encouraging household energy conservation Ehrhardt-Martinez et al. (2010), and for electric utilities aiming to understand air conditioner adoption patterns, and determinants of electricity demand more broadly.

In parallel, a growing number of econometric studies use smartmeter data or monthly billing data to evaluate energy efficiency programs ex-post Boomhower and Davis (2016); Fowlie et al. (2015); Meyer et al. (2017). Such studies are an indispensable component of efforts to realize high levels of cost-effective energy efficiency required to mitigate the effects of climate change. These econometric studies, particularly those that attempt to understand changes in electricity consumption patterns throughout the day, would ideally control for the presence of major time-dependent loads such as air conditioning.

Unfortunately, utility data generally do not report such household-level characteristics. There is as yet no established method of determining the presence of air conditioning within a household using widely available electricity consumption and ambient temperature data. Existing NILM techniques generally use sub-hourly electricity consumption data, often at 1-minute or even 1-second intervals. In the econometric literature, Boomhower & Davis attempt to verify the presence of residential air conditioning through inspection of electricity consumption data over time Boomhower and Davis (2016).

We apply standard machine learning techniques to identify the presence of air conditioning in households using monthly electricity consumption data, which are available to any US utility and many academic researchers, together with publicly available temperature data. In addition, we cluster households on the basis of monthly electricity consumption to see whether households with air conditioning or other distinctive features coincide with the learned clusters.

In addition to assisting with econometric studies of electricity consumption, a reliable air conditioner identification algorithm can help utilities improve targeting in program outreach, e.g. for air conditioner-specific demand-side management programs.

2 Background

For more than thirty years, researchers have been attempting to use household electricity consumption data to learn about the presence of electrical appliances within households, and the corresponding breakdown of household electricity consumption. This practice is called non-intrusive load monitoring (NILM). Much of this work focuses on electricity consumption data with sub-hourly resolution, with measurements every fifteen minutes, every minute, and in some cases every second or less Wytock and Kolter (2014); Giri and Berges (2015); Drenker and Kader (1999); Cominola et al. (2017); Lin et al. (2016); Hart (1985); Esa et al. (2016); Anderson et al. (2011); Kalluri et al. (2016); Chahine et al. (2011); Farinaccio and Zmeureanu (1999); Figueiredo et al. (2012); Amenta and Tina (2015); Tsai and Lin (2012); Batra et al. (2014); Zeifman and Roth (2011); Biansoongnern and Plungklang (2016); Perez et al. (2014); Norford and Leeb (1996); Marceau and Zmeureanu (2000); Su et al. (2016); Rahimpour et al. (2015); Parson et al. (2012); Aladesanmi and Folly (2015). Many of these papers explicitly estimate the presence of air conditioning with high accuracy Drenker and Kader (1999); Norford and Leeb (1996); Marceau and Zmeureanu (2000). Others attempt to diaggregate air conditioner electricity consumption, often implicitly identifying the presence of an air conditioner Wytock and Kolter (2014); Su et al. (2016); Aladesanmi and Folly (2015); Biansoongnern and Plungklang (2016); Zeifman and Roth (2011); Batra et al. (2014). However, none do so using data with coarser than hourly resolution, let alone monthly resolution. In the building energy efficiency literature, estimating building parameters, such as thermostat set-points, is common practice Fels (1986); Enriquez et al. (2017); Smullin (2016). Enriquez et al. (2017) estimate building-level thermal parameters, such as building air exchange rates, using minute-resolution data. Fels (1986) estimates thermostat set points and the sensitivity of a building's electricity consumption to temperature using monthly electricity consumption data, with Smullin (2016) doing the same with minute-resolution data. However, neither attempt to learn the presence of air conditioning in a household with out-of-sample predictive methods.

We attempt to combine these literatures, predicting the presence of air conditioning from monthly electricity consumption and monthly average temperature data using three methods. First, we compare estimated building sensitivity to high temperature values using the regression. Next, we apply a random forest classifier to monthly electricity consumption. Finally, we cluster monthly electricity consumption profiles to determine to what extent these learned clusters correspond to several household features.

3 Problem statement

Electric utilities and energy researchers seeking to understand determinants of electricity consumption often have access to residential electricity consumption data, at least at a monthly level, but do not have data on the presence of air conditioning, an important factor in such analyses. We develop a method to predict the presence of air conditioning in a household using monthly residential electricity consumption data and publicly available temperature data. We hypothesize that it is possible to classify such households with at least 95% accuracy out-of-sample.

4 Data

We use residential electricity consumption data from Pecan Street Inc., which provides electricity consumption information for 1,429 households, including appliance-level electricity consumption measurements for 690 households, primarily in and around Austin, Texas. Appliance-level data are collected using devices installed on each monitored appliance. The data are then transmitted remotely back to Pecan Street Inc. Data are available at www.pecanstreet.org.

Participants are recruited on a voluntary basis by Pecan Street Inc. Data collection began at some households in 2012, with additional households recruited on an ongoing basis. We use electricity data through December 2015. The sample is fairly wealthy, with a mean household income of $157,000, and a standard deviation of $145,000 Glasgo et al. (2016). Although the sample itself may not be representative of typical households in Texas, or anywhere else, we expect that signatures of the presence of air conditioning in this sample likely have external validity, at least to households with air conditioning in similar climates within the United States.

The dataset contains information regarding the presence of air conditioning, as well as many other appliances, along with other household-level features, such as the type of building, square footage, year of home construction, and participation in energy-related programs.

We aggregate these data to monthly electricity consumption to match the resolution available to electric utilities without smartmeters.

The Pecan Street dataset includes households which are not submetered, and thus may have air conditioning even if the data do not say so explicitly. To account for this, we include only households that have submetered electricity consumption data for at least one non-air conditioning appliance. We include only households with at least 12 months of valid data. We exclude households listed as containing air conditioning that do not have any recorded air conditioner use. This leaves 482 valid households, 449 with central or window air conditioning, and 33 without AC.

For this subset of the data, mean monthly electricity consumption for all households is 856 kWh, with a 5th percentile of 301 kWh and a 95th percentile of 1,783 kWh. Of this, electricity consumption from the air conditioners themselves has a mean of 240 kWh, with a 5th percentile of 0 kWh and a 95th percentile of 562 kWh. This whole-year average is likely a low estimate of air conditioner electricity consumption in summer months. For households with air conditioning, mean monthly electricity consumption is 916 kWh, with a 5th percentile of 315 kWh, and a 95th percentile of 7,831 kWh. For households without air conditioning, mean monthly electricity consumption is 456 kWh, with a 5th percentile of 207 kWh, and a 95th percentile of 1090 kWh.

Within our sample, the mean household with air conditioning has 27 months of data, with a minimum of 12 months, and a maximum of 48 months. The mean household without air conditioning has 21 months of data, with a minimum of 14 months, and a maximum of 41 months.

We believe all households labeled as containing air conditioning are correctly labeled, as the dataset includes submetered air conditioner electricity consumption data. We are not convinced that all households labeled as not containing air conditioning do not in fact contain air conditioning. Electricity consumption in some households labeled as non-AC increases sharply in warmer months in a way that we believe is unlikely in the absence of air conditioning, which tends to be by far the largest heat-sensitive load.

We add temperature data collected from the KATT weather station in Austin, Texas. These data are available at www.wunderground.com/history. These data include daily high, low, and average temperature, as well as humidity and a number of other metrics. We use average monthly temperature in our analysis.

5 Methods

We perform three classes of analysis. The first attempts to classify households with and without air conditioning using an engineering-inspired regression of electricity consumption v. temperature. The second applies the random forest to a version of the same data. The third clusters households on the basis of electricity consumption data, and tests the descriptive capabilities of these clusters for several household characteristics, such as the presence of air conditioning or solar photovoltaics. All three analyses draw from the 482 households from in and around Austin, Texas. We remove all monthly electricity consumption measurements of less than 2.5 kWh, roughly equivalent to electricity consumption from a single 14W compact fluorescent lightbulb over one week.

We randomly divide the data into training, validation, and test sets. We stratify the randomization, separately allocating households with and without air conditioning. We place one third of the households with air conditioning in the training (149 households), validation (149 households), and test (151 households) datasets. For households without air conditioning, we place 3/7 of households (15 households) in the training set, with 2/7 each in the validation (9 households) and test (9 households) datasets. We incorporate more non-AC households into the training dataset to improve training capabilities, leaving enough non-AC households in the validation and test datasets to produce meaningful out-of-sample prediction results.

Because non-AC households are such a small fraction of the sample, training on based on simple classification accuracy would place a much higher penalty on the false negative rate, AC homes classified as non-AC, than the false positive rate, non-AC homes classified as AC. We address this by optimizing the negative predictive value (NPV) of our classifiers, where NPV is the ratio of correctly-classified non-AC households to total non-AC classifications. We constrain this optimization to ensure that at least 50% of non-AC households are correctly classified. This ensures that the resulting classifiers correctly classify a substantial fraction of non-AC households with meaningful predictive power.

We train all classifiers on the training dataset, and test preliminary out-of-sample classification accuracy on the validation dataset. To ensure our estimates do not overfit to the training data, we test out-of-sample classification accuracy on the test set only after finalizing the form and parameter values of a classifier.

Our engineering-inspired analysis takes advantage of the positive correlation between ambient temperature and air conditioner electricity consumption, commonly estimated to begin at 65 degees Fahrenheit Eto (1985). For each household, we fit a line to monthly electricity consumption as a function of mean monthly ambient temperature for all months with mean temperature above 65 degrees. The regression has the following form for each household:

$$kWh_m = \beta t_m + \epsilon_m \tag{1}$$

Where kWh_m is monthly household electricity consumption in month m, measured in kWh. t_m is mean monthly temperature in month m, in degrees Fahrenheit. ϵ_m is a random error term, assumed to have mean zero. A month, m, is only included in the regression if $t_m \geq 65°F$.

We then predict the presence of air conditioning based on the slope of this fitted line, β. We select a classification threshold, γ to maximize NPV while correctly classifying at least 50% of households without air conditioning. We then test prediction accuracy on the test dataset.

Our second classification method selects the final twelve months of electricity consumption for each household, and treats electricity consumption in each month as a separate feature. We include only households for which the final twelve months of data have no missing months. As a result, the training dataset contains 155 households, 140 with air conditioning and 15 without; the validation dataset contains 152 households, 143 with air conditioning and 9 without; the test dataset contains 150 households, 141 with air conditioning and 9 without.

For each household, we normalize by average electricity consumption over the 12-month period to control for differences in the overall level of electricity consumption between households. Because all households in this sample come from the same city, weather is essentially the same across the sample, although not all households use the same twelve months. We then fit a random forest classifier to the training households using the `randomForest` package in R, using the default settings for Breiman's classification algorithm Liaw and Wiener (2002); Breiman (1999). We select the random forest due to its flexibility in learning complex functional forms.

The likely presence of noise in our labels adds another layer of difficulty to this classification task, which already has a small set of households without air conditioning. We expect that this will pose particular difficulties for classical machine learning methods, such as the random forest, which risks overfitting to noise.

In the clustering analysis, we use the same dataset as in the random forest analysis, and apply the Hartigan-Wong k-means clustering algorithm using an squared-error distance metric R Core Team (2013); Hartigan and Wong (1979), selecting the lowest total within-cluster sum of squares from 1000 iterations. We select the number of clusters by maximizing NPV for the presence of air conditioning for at least 50% of households. We use both the training and test datasets for training in this analysis. We then tabulate overlap between a household's assigned cluster, and the presence of air conditioning and other appliances, as well as other household-level features, such as the square footage, year of home construction, and the presence of solar photovoltaics. We then check whether such a clustering analysis points to informative underlying predictive patterns in the data, which can be exploited for future analysis.

Finally, we compare the above AC identification techniques with human classification using the same electricity consumption and temperature data used in the linear model. A human classifies each household based on a plot of average electricity consumption v. monthly average temperature, similar to Figure 1. The human knows the number of households with and without AC in advance, but households are displayed in random order, and the human does not see these labels, or the household identification numbers during the classification tasks. Given a dataset, and the number of non-AC households, N, the human selects the N households they believe most likely do not contain AC. We term classifications based on this set as "Fixed Number". The human also selects a larger subset of households that they believe may also not contain AC. We term classifications based on this set as "Ambiguous→no AC", because all ambiguous cases are classified as non-AC. The human completes this classification task for the training, validation, and test datasets in that order, evaluating predictive performance after classifying each of the three datasets.

When assigning households to the training, validation, and test sets, we use an R random seed of 100, in version 1.0.136 of RStudio, and version 3.3.2 of R.

6 Analysis

In our engineering-inspired regression analysis, we classify households as containing an air conditioner if their slope coefficient, β, from equation 1 is greater than or equal to a threshold value, γ. We determine γ from the training data, selecting a value that maximizes negative predictive value (NPV), the number of correctly-predicted non-AC households as a fraction of the total number classified as non-AC, for at least half of the 15 non-AC households in the sample. We consider thresholds in increments of $0.25kWh/°F$. We test the classification accuracy on the validation dataset, which was also used for exploratory analysis. After the model has been finalized, we test classification accuracy on the test dataset.

We then apply the random forest classification algorithm described in the Methods section to the training dataset, considering the final twelve months of data for all households without missing electricity consumption data during their final twelve months. We consider twelve features, representing household electricity consumption in each month of the year, divided by mean household electricity consumption over the twelve-month period considered. As in the linear analysis, we test the classification accuracy on the training dataset, which was also used for exploratory analysis. After the model has been finalized, we test classification accuracy on the test dataset.

Finally, we cluster the data using the k-means algorithm described in the Methods section. We select the number of clusters that produces a classifier that maximizes NPV for at least half of the households without air conditioning. Such a classifier classifies households as AC if they are in one set of clusters, and non-AC if not. We constrain this classifier to correctly classify at least half of all non-AC households. We train on the combined training and validation datasets, as this method is not primarily aimed at producing out-of-sample predictions. We then assign households within the test dataset to the nearest of these clusters using a squared-error distance metric, and test the out-of-sample classification positive predictive value (PPV) and NPV of the clustering-based classifier. PPV is the number of correctly-classified AC households divided by the total number classified as AC.

We also test the classification performance of the same clusters for several other household-level features, the presence of rooftop photovoltaics, an electric car charging port, a pool pump, a year of construction greater than or equal to the in-sample median year of 2007, and household square-footage greater than the in-sample median of 1720 sq. ft. We use the combined training and test data to compute classification accuracy. Many of these quantities have far more false than true values, or vice versa. For this reason, we use the same classification metric as for air conditioning. For whichever condition has fewer instances, either the true or false case, we select the clusters that maximize predictive value for at least 50% of instances of that case. If there are more negatives than positives, we maximize PPV. If there are more positives, we maximize NPV. By construction, these clusters will perform at least as well as chance in-sample.

7 Results

7.1 Linear method

In our engineering-inspired regression, equation 1, we find that during warmer months, with average temperature above $65°F$, monthly electricity consumption in households in the training dataset increases by an average of $39kWh/°F$. Following equation 1, we use β to denote this household-level quantity, the increase in monthly electricity consumption per $°F$ increase in temperature. The standard deviation is $33kWh/°F$, with 25th and 75th percentile values of $13kWh/°F$ and $53kWh/°F$ respectively, and maximum and minimum values of $277kWh/°F$ and $-7kWh/°F$. For households with AC, the mean is $42kWh/°F$, with a standard deviation of $32kWh/°F$. For non-AC households, the distribution is skewed, with a median of $2/°F$, but a mean of $7kWh/°F$, a standard deviation of $20kWh/°F$, and a maximum of $78kWh/°F$. Figure 1 shows an example of these fitted lines, and the underlying electricity consumption and temperature data, for example households with and without AC.

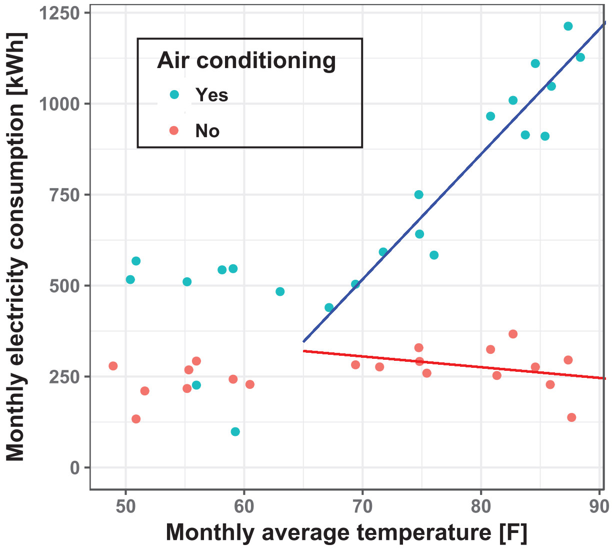

Figure 1: Monthly household electricity consumption v. monthly average temperature for one household with air conditioning (AC), blue, and one without, red. The blue and red lines are fitted to all months for each household with an average temperature of at least $65°F$.

We attempt to classify households by thresholding on γ. We find that a threshold of $4.5kWh/°F$ maximizes negative predictive value (NPV), the number of correctly classified households without air conditioning divided by the total number of households classified as not having AC, while correctly classifying at least 50% of households without air conditioning. Here, in-sample NPV is 71%, correctly classifying 12 of 15 households without AC. Positive predictive value (PPV) is 98%, correctly classifying 144 of 149 households with AC. The confusion matrix of this predictor is shown in Table 1. Recall that within the training data, there are 149 AC households, or 91% of the total, and 15 non-AC households, or 9% of the total.

Among the test households shown in Table 1, we see a decline in predictive performance, with PPV at 96%, in part due to the smaller number of non-AC households, and NPV declining to 50%, with 6 of 9 non-AC households misclassified. Regardless, a decline in predictive performance on the test dataset is expected, as the threshold parameter, γ, was optimized for the training dataset, and thus runs a risk of overfitting.

After selecting the final model, parameters, and data format using the training and test datasets, we test the classification accuracy of this linear method on the test dataset, which had not been used for any prior analysis. We find similar results to the training data, with a PPV of 97%, and an NPV of 67%, correctly classifying 4 of 9 non-AC households, and 149 of 151 AC households. In Table 1, we show the resulting confusion matrix.

Table 1: Linear method predicted (rows) v. labelled (columns) presence of air conditioning in the training, validation, and test households. Quotes denote predictions. Bottom row is positive predictive value (PPV) for AC households, and negative predictive value (NPV) for non-AC households.

Training	AC	no AC	Validation	AC	no AC	Test	AC	no AC
"AC"	144	3	"AC"	146	6	"AC"	149	5
"no AC"	5	12	"no AC"	3	3	"no AC"	2	4
% Correct	97%	80%	% Correct	98%	33%	% Correct	99%	44%
PPV/NPV	98%	71%	PPV/NPV	96%	50%	PPV/NPV	97%	67%

7.2 Random forest

The random forest predicts perfectly in-sample on the training data, as shown in Table 2. Prediction on the validation dataset, also shown in Table 2, has a PPV of 100%, meaning that all AC households correctly identified, and an NPV of 33%. The test dataset has a PPV of 97%, and an NPV of 71%, with five of nine households without AC correctly classified.

The two most important months selected by the random forest, using a Gini index metric, are July and November, indicating, unsurprisingly, that the random forest is comparing electricity consumption in Summer and Fall.

Table 2: Random forest predicted (rows) v. labelled (columns) presence of air conditioning in the training, validation, and test households. Quotes denote predictions. Bottom row is positive predictive value (PPV) for AC households, and negative predictive value (NPV) for non-AC households.

Training	AC	no AC	Validation	AC	no AC	Test	AC	no AC
"AC"	140	0	"AC"	143	6	"AC"	139	4
"no AC"	0	15	"no AC"	0	3	"no AC"	2	5
% Correct	100%	100%	% Correct	100%	33%	% Correct	99%	56%
PPV/NPV	100%	100%	PPV/NPV	96%	100%	PPV/NPV	97%	71%

7.3 Clustering

In the clustering analysis, we find that a setup with 22 clusters maximizes NPV for at least half of all non-AC households at 62%. Figure 2 shows normalized electricity consumption for the 22 cluster centers, and the breakdown within each cluster of households with and without AC.

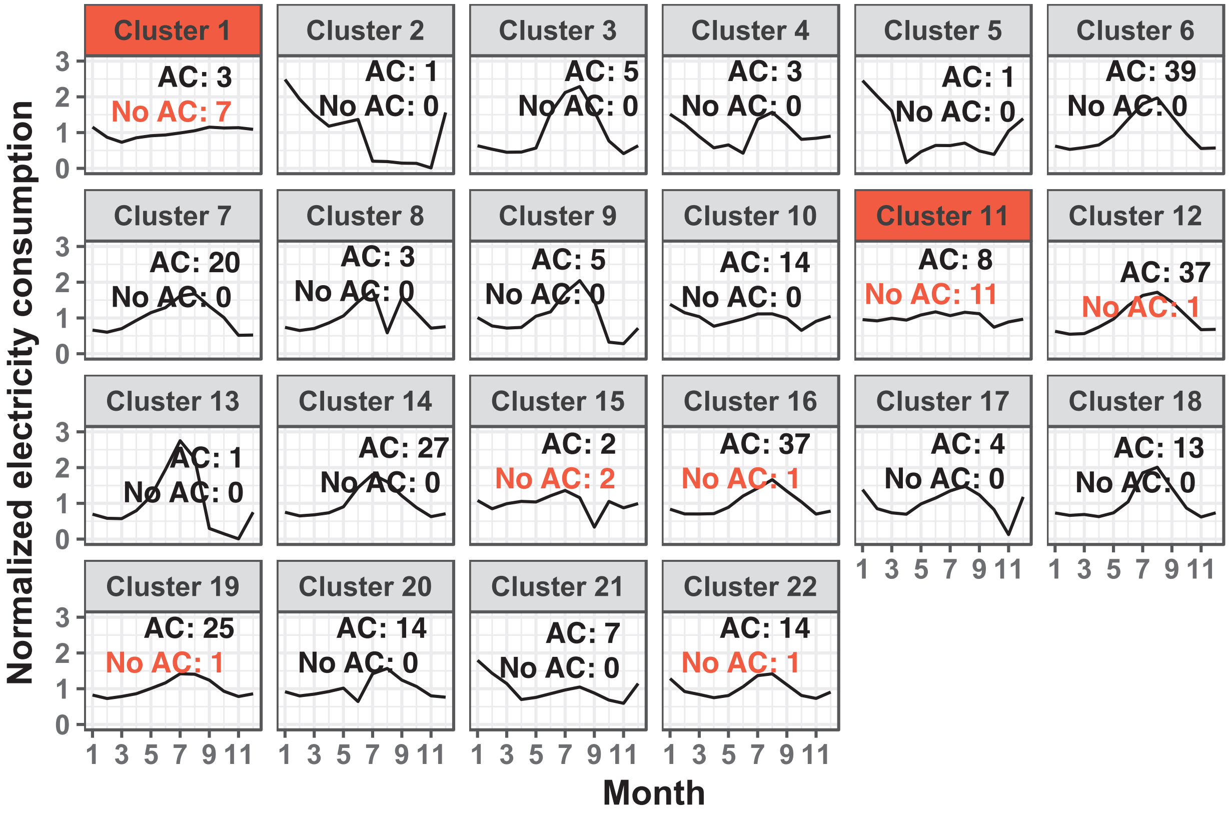

Figure 2: Learned clusters of monthly household electricity consumption, with counts of the number of households in each cluster with and without air conditioning (AC). The relatively flat Clusters 1 and 11, highlighted in red, contain the majority of households without AC. Most households with AC have a pronounced summer peak. All households without AC are highlighted in red.

Clusters 1 and 11, which have essentially flat electricity consumption throughout the year, contain 18 of 24 households without AC, as well as 11 of 283 with AC. Clusters 3, 6, 7, 9, 12, 13, 14, 16, 18, 19, and 20 all have pronounced summer peaks in electricity consumption, and collectively contain 227 of 307 households with AC, and only 3 of 24 households without AC. Clusters 4, 10, 17, 21, and 22, which have summer peaks alongside pronounced winter peaks, collectively contain 42 more households with air conditioning and 1 without. Clusters 2, 5, 8, and 15 have more unusual shapes, with sudden jumps in monthly electricity consumption. These clusters collectively contain 9 households with air conditioning, and 2 without. Overall, a classification strategy that classifies households in Clusters 1 and 11 in as non-AC, and households in all other clusters as AC has an in-sample PPV of 98%, and an NPV of 62%.

We assign households in the test dataset to the fitted clusters using an L2 distance metric, and test the out-of-sample prediction accuracy of this clustering-based classification method. The test dataset, was not used during model training, testing, or model selection. In Table 3, we show that classifying all households in Clusters 1 and 11 as non-AC, and all other households as AC, yields a PPV of 98%, and an NPV of 40%, with a base rate of 94% AC and 6% non-AC.

Table 3: Predicted (rows) v. labelled (columns) presence of air conditioning in the combined training and validation (in-sample), test households (out-of-sample), using a clustering-based classifier. We classify all households in Clusters 1 and 11 as non-AC, and all others as AC. Quotes denote predictions. Bottom row is positive predictive value (PPV) for AC households, and negative predictive value (NPV) for non-AC households.

Training + Test	AC	no AC	Test	AC	no AC
"AC"	272	6	"AC"	132	3
"no AC"	11	18	"no AC"	9	6
% Correct	96%	75%	% Correct	94%	67%
PPV/NPV	98%	62%	PPV/NPV	98%	40%

We use the same clusters to classify households based on other features. The best-predicted feature is the presence of photovoltaics, with a PPV of 53% and an NPV of 72%, with an underlying breakdown of 38% of households with photovoltaics, and 62% without. Predictions of household area greater than or equal to the median have a PPV of 58% and an NPV of 61%, with an underlying breakdown of 49% of households greater than the median size, and 51% smaller. Predictions of the presence of an electric car charging port have a PPV of 35% and an NPV of 91%, with an underlying breakdown of 84% with, and 16% without. For pool pumps, PPV is 20% and NPV is 95%, with an underlying breakdown of 94% of households without pool pumps, and 6% with.

The author completed the human identification task, described in the Methods section. Tables 4 and 5 show performance for "Fixed Number" and "Ambiguous→no AC" human classification respectively. Human classification performs similarly, but slightly worse than the linear and random forest techniques in essentially all cases, and by all metrics.

Table 4: Human "Fixed Number" classification, predicted (rows) v. labelled (columns) presence of air conditioning in the training, validation, and test households. Quotes denote predictions. Bottom row is positive predictive value (PPV) for AC households, and negative predictive value (NPV) for non-AC households.

Training	AC	no AC	Validation	AC	no AC	Test	AC	no AC
"AC"	143	6	"AC"	145	4	"AC"	146	5
"no AC"	6	9	"no AC"	4	5	"no AC"	5	4
% Correct	96%	60%	% Correct	97%	56%	% Correct	97%	44%
PPV/NPV	96%	60%	PPV/NPV	97%	56%	PPV/NPV	97%	44%

Table 5: Human "Ambiguous→no AC" predicted (rows) v. labelled (columns) presence of air conditioning in the training, validation, and test households. Quotes denote predictions. Bottom row is positive predictive value (PPV) for AC households, and negative predictive value (NPV) for non-AC households.

Training	AC	no AC	Validation	AC	no AC	Test	AC	no AC
"AC"	138	2	"AC"	141	3	"AC"	144	4
"no AC"	11	13	"no AC"	8	6	"no AC"	7	5
% Correct	92%	87%	% Correct	95%	67%	% Correct	95%	56%
PPV/NPV	99%	54%	PPV/NPV	98%	43%	PPV/NPV	97%	42%

8 Discussion

We find that monthly household electricity consumption and outdoor temperature data are useful predictors of the presence of air conditioning in a household. Both the quantitative and human predictors demonstrate there is indeed a signal in the data that is predictive of the presence of air conditioning. Both the linear and random forest methods have out-of-sample accuracy greater than the 95% level hypothesized. However, overall accuracy allows a high positive predictive value, close to 99%, to mask a substantially lower negative predictive value, closer to 50%. Based on these results, these methods show promise, but likely require further analysis with larger datasets, ideally with more certainty in the non-AC labels, before these methods can be used in econometric applications, such as analysis of air conditioner adoption behavior, which would likely require closer to 95% positive and negative predictive value.

The clustering analysis has lower predictive performance than the linear and random forest methods in terms of overall accuracy, and positive and negative predictive value. This analysis does, however, illustrate that households without AC overwhelmingly tend to have relatively flat electricity consumption throughout the year, at least within our sample. The resulting clusters do not produce particularly informative predictions of other household-level quantities, such as the presence of solar photovoltaics. This is unsurprising, given that air conditioning is both a particularly large, and particularly seasonal consumer of electricity, which makes it more easily discernible from monthly data.

9 Limitations

The Pecan Street dataset is an indispensable resource for research into energy consumption behavior. Still, the data introduce several limitations into our analysis. First, the sample is relatively small, consisting of less than 500 useable households, over 90% of which have air conditioning. This small sample size poses challenges both for training statistical learners, and reliably testing their performance.

The households are overwhelmingly from a single metropolitan area, Austin, Texas, and are collected on a voluntary basis, resulting in numerous selection effects, including relatively high income. Only some of these selection effects are, or even can be measured and controlled for. This introduces threats to external validity in different regions and for different subpopulations.

The presence of appliance-level electricity consumption measurements gives us confidence in the labels that indicate the presence of air conditioning. There is more room for error in non-AC household labels, as a household may have air conditioning that simply has not been noted in the metadata. We suspect this to be the case for roughly 10% of households labeled "non-AC" in our sample, as 4 of 33 have temperature v. electricity consumption lines with slopes above $40kWh/^{\circ}F$, above the median value of $39kWh/^{\circ}F$ for households with air conditioning. This could be explained by the presence of another large, temperature sensitive load, but we do not have a plausible hypothesis of what such a device would be, other than some form of air conditioning.
Even with perfectly labeled data, numerous behavioral and location-dependent factors introduce threats to internal and external validity. Summer vacations in the warmest months of summer, with a corresponding drop in electricity consumption, will have high leverage in the linear model, potentially masking the presence of air conditioning. The random forest, while more flexible in form, has the potential to rely on idiosyncratic predictors of the presence of air conditioning that may not translate well to other regions or population segments. For example, a predictor based on a June-August peak in electricity consumption would likely fail in a southern hemisphere contact, where summer occurs in January-March.

10 Conclusions

These results demonstrate that monthly household electricity consumption is a strong predictor of the presence of air conditioning, at least among relatively affluent households in Austin, Texas. Of the three classifiers tested, the linear model likely has the strongest external validity, as it is explicitly based on the well-understood relationship between outdoor temperature and air conditioner use. The random forest method likely has the most predictive potential of the three methods, due to its ability to learn complex relationships, but would likely need to be re-trained before it can be applied in a new climate. The clustering model provides useful insights into the electricity consumption profiles associated with air conditioning, but may have less potential as a classifier in its own right. Overall, this line of inquiry shows promise for econometricians seeking to understand the determinants of residential electricity consumption behavior, particularly adoption of air conditioning. However, improvements in negative predictive value, the ability to predict the absence of air conditioners, are likely necessary before these methods can be incorporated into such applications. Such improvements will likely require additional data, ideally from areas with different climates. In addition to econometric applications, an accurate air conditioner identification algorithm could help electric utilities better forecast changes in electricity demand growth and peak electricity demand, thus informing long-term infrastructure investment decisions, and energy efficiency and demand-side management program design.

11 Acknowledgments

I gratefully acknowledge the Pecan Street Inc. for providing data access. I also gratefully acknowledge my research advisors, J. Zico Kolter and Inês Azevedo. I acknowledge and thank Brock

Glasgo and Nichole Hanus for introducing me to important building science literature and techniques. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1252522. This work was funded in part by the Center for Climate and Energy Decision Making (SES-0949710 and SES-1463492), through a cooperative agreement between the National Science Foundation and Carnegie Mellon University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Aladesanmi, E. J. and Folly, K. A. (2015). Overview of non-intrusive load monitoring and identification techniques. *IFAC-PapersOnLine*, 48(30):415–420.

Amenta, V. and Tina, G. M. (2015). Load Demand Disaggregation Based on Simple Load Signature and User's Feedback. *Energy Procedia*, 83:380–388.

Anderson, K., Ocneanu, A. F., Benitez, D., Carlson, D., Rowe, A., and Berges, M. (2011). BLUED: A fully labeled public dataset for event-based non-intrusive load monitoring research. In *2nd Workshop on Data Mining Applications in Sustainability (SustKDD)*, page 2012.

Baker, M., Fuller, M., and Hicks, G. (2016). Can We Find the End Use in Smart Metering Data? *2016 ACEEE Summer Study on Energy Efficiency in Buildings*.

Batra, N., Kelly, J., Parson, O., Dutta, H., Knottenbelt, W., Rogers, A., Singh, A., and Srivastava, M. (2014). NILMTK: an open source toolkit for non-intrusive load monitoring. pages 265–276. ACM Press.

Biansoongnern, S. and Plungklang, B. (2016). Non-Intrusive Appliances Load Monitoring (NILM) for Energy Conservation in Household with Low Sampling Rate. *Procedia Computer Science*, 86:172–175.

Boomhower, J. and Davis, L. W. (2016). Do Energy Efficiency Investments Deliver at the Right Time? *Working paper*.

Breiman, L. (1999). Random forests. *UC Berkeley TR567*.

Chahine, K., Drissi, K. E. K., Pasquier, C., Kerroum, K., Faure, C., Jouannet, T., and Michou, M. (2011). Electric Load Disaggregation in Smart Metering Using a Novel Feature Extraction Method and Supervised Classification. *Energy Procedia*, 6:627–632.

Cominola, A., Giuliani, M., Piga, D., Castelletti, A., and Rizzoli, A. (2017). A Hybrid Signature-based Iterative Disaggregation algorithm for Non-Intrusive Load Monitoring. *Applied Energy*, 185:331–344.

Drenker, S. and Kader, A. (1999). Nonintrusive monitoring of electric loads.

Ehrhardt-Martinez, K., Donnelly, K. A., and Laitner, J. A. . (2010). Advanced Metering Initiatives and Residential Feedback Programs: A Meta-Review for Household Electricity-Saving Opportunities. Technical report, American Council for an Energy-Efficient Economy, Washington, DC.

EIA (2016). Electric Power Annual 2015. Technical report, Energy Information Administration, Washington, DC.

Enriquez, R., Jimenez, M., and Heras, M. (2017). Towards non-intrusive thermal load Monitoring of buildings: BES calibration. *Applied Energy*, 191:44–54.

EPA (2016). Inventory of U.S. Greenhouse Gas Emissions and Sinks: 1990 2014. Technical Report EPA 430-R-16-002, U.S. Environmental Protection Agency.

Esa, N. F., Abdullah, M. P., and Hassan, M. Y. (2016). A review disaggregation method in Non-intrusive Appliance Load Monitoring. *Renewable and Sustainable Energy Reviews*, 66:163–173.

Eto, J. (1985). Characterizing the effects of weather on commercial building energy use. In *Building energy simulation conference, Seattle, WA*.

Farinaccio, L. and Zmeureanu, R. (1999). Using a pattern recognition approach to disaggregate the total electricity consumption in a house into the major end-uses. *Energy and Buildings*, 30(4):245–259.

Fels, M. F. (1986). PRISM: an introduction. *Energy and Buildings*, 9(1-2):5–18.

Figueiredo, M., de Almeida, A., and Ribeiro, B. (2012). Home electrical signal disaggregation for non-intrusive load monitoring (NILM) systems. *Neurocomputing*, 96:66–73.

Fowlie, M., Greenstone, M., and Wolfram, C. (2015). Do energy efficiency investments deliver? Evidence from the weatherization assistance program. Technical report, National Bureau of Economic Research.

Giri, S. and Berges, M. (2015). An energy estimation framework for event-based methods in Non-Intrusive Load Monitoring. *Energy Conversion and Management*, 90:488–498.

Glasgo, B., Azevedo, I. L., and Hendrickson, C. (2016). How much electricity can we save by using direct current circuits in homes? Understanding the potential for electricity savings and assessing feasibility of a transition towards DC powered buildings. *Applied Energy*, 180:66–75.

Hart, G. W. (1985). Prototype nonintrusive appliance load monitor. Technical report, Massachussets Institute of Technology, Electric Power Research Institute, Concord, Massachussetts.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics*, 28(1):100.

Kalluri, B., Kamilaris, A., Kondepudi, S., Kua, H. W., and Tham, K. W. (2016). Applicability of using time series subsequences to study office plug load appliances. *Energy and Buildings*, 127:399–410.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

Lin, S., Zhao, L., Li, F., Liu, Q., Li, D., and Fu, Y. (2016). A nonintrusive load identification method for residential applications based on quadratic programming. *Electric Power Systems Research*, 133:241–248.

Marceau, M. L. and Zmeureanu, R. (2000). Nonintrusive load disaggregation computer program to estimate the energy consumption of major end uses in residential buildings. *Energy Conversion and Management*, 41(13):1389–1403.

Meyer, R. M., Sherwin, E. D., and Azevedo, I. L. (2017). Unintended consequences of California energy efficiency rebate programs: without recycling, rebound. *Working paper.*

Norford, L. K. and Leeb, S. B. (1996). Non-intrusive electrical load monitoring in commercial buildings based on steady-stateandtransientload-detection algorithms. *Energy and Buildings*, 24:51–64.

Parson, O., Ghosh, S., Weal, M., and Rogers, A. (2012). Non-Intrusive Load Monitoring Using Prior Models of General Appliance Types. In *AAAi*.

Perez, K. X., Cole, W. J., Rhodes, J. D., Ondeck, A., Webber, M., Baldea, M., and Edgar, T. F. (2014). Nonintrusive disaggregation of residential air-conditioning loads from sub-hourly smart meter data. *Energy and Buildings*, 81:316–325.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rahimpour, A., Qi, H., Fugate, D., and Kuruganti, T. (2015). Non-intrusive load monitoring of HVAC components using signal unmixing. In *Signal and Information Processing (GlobalSIP), 2015 IEEE Global Conference on*, pages 1012–1016. IEEE.

Smullin, S. J. (2016). Thermostat metrics derived from HVAC cycling data for targeted utility efficiency programs. *Energy and Buildings*, 117:176–184.

Su, S., Yan, Y., Lu, H., Kangping, L., Yujing, S., Fei, W., Liming, L., and Hui, R. (2016). Non-intrusive load monitoring of air conditioning using low-resolution smart meter data. In *Power System Technology (POWERCON), 2016 IEEE International Conference on*, pages 1–5. IEEE.

Tsai, M.-S. and Lin, Y.-H. (2012). Modern development of an Adaptive Non-Intrusive Appliance Load Monitoring system in electricity energy conservation. *Applied Energy*, 96:55–73.

Wytock, M. and Kolter, J. Z. (2014). Contextually Supervised Source Separation with Application to Energy Disaggregation. *Association for the Advancement of Artificial Intelligence.*

Zeifman, M. and Roth, K. (2011). Nonintrusive appliance load monitoring: Review and outlook. *IEEE transactions on Consumer Electronics*, 57(1).