2015

An Exploratory Analysis of Electric Usage Data from the Vermont Energy Investment Corporation

Colby John Nadeau
University of Vermont

An Exploratory Analysis of Electric Usage Data from the Vermont Energy Investment Corporation

An Honors Thesis by:
Colby Nadeau

Advisor: Richard Single, Ph.D.

Table of Contents

Table of Figures

INTRODUCTION

This paper examines energy usage patterns of residential and commercial households in the state of Vermont. There are three main objectives of this paper. The first is to aggregate the data from the "high bill" calls that the Vermont Energy Investment Corporation (VEIC) receives. This is accomplished by cleaning and organizing the data into one Excel spreadsheet using Google Refine. The second objective is to predict the presence of an appliance based solely on the total usage of a home. This is accomplished using data collected from 24 high performance homes in Vermont. This data set is analyzed using histograms, correlations, clustering, random forests, tree plots, and heat maps. The final objective is to predict the future electric usage for the commercial accounts in Vermont. This is accomplished by analyzing six different models on the monthly usage from the commercial accounts and then comparing the results of the models to determine the best approach.

PRELIMINARY WORK

To accomplish the objectives of this paper some preliminary work needed to be performed. An understanding about regular expressions needed to be developed, and to better comprehend regular expressions I watched a YouTube R course given by Roger Peng. The YouTube course contained videos on regular expressions as well as videos on the various apply() functions in R. Regular expressions became an important part of this thesis as they were used in every major project. I created R-scripts with some practice data to become more familiar with R and some of its functions. An older script was used as the basis for most of this work. Using the Iris, Kyphosis, and Cars data from R, as

well as some loan information downloaded from the internet, these new scripts experimented with R's apply() functions, kmeans clustering, and tree plots. This helped to develop a much better comprehension of the apply() functions which also became an integral aspect of this thesis.

METHODS

DATA AGGREGATION FOR "HIGH BILL" CALLS

VEIC receives "high bill" calls to their customer service department. A "high bill" call occurs when a customer obtains their electric bill and believes that it is too high. They then call VEIC to discuss possible reasons why the bill was higher than they expected. Each of the different customer service representatives puts the relevant data from the call into an excel file and saves it for future follow up information to the customer. The problem is that the high bill calls had been handled by many different people and therefore were not all in the same format. VEIC had over 200 of these "high bill" calls to investigate. Working through each of the "high bill" calls required finding the relevant information from each of the excel files and copying it into one large Excel spreadsheet. This spreadsheet included the player name (i.e., customer ID), appliance, quantity, hours per day, hours per month, wattage, kWh per month, date of call, a comment, and the default values for hours per day, hours per month, and kWh per month when available. The comment field listed which appliances were asked about, even though they were out of season (winter and summer appliances), as well as any other pertinent information. Most of the "high bill" call files did not include the account number for the player name, so KITT was used to find the applicable account number based on whether or not a "high

bill" call was made. KITT is the database that VEIC uses to store all of the information that they obtain for the accounts in Vermont. KITT stores all of the identification information, as well as the monthly electric usage readouts for all accounts. It is also used to keep track of which accounts have contacted VEIC, either with questions or to undergo a project with them. If there was no account number, or the correct player could not be found in KITT, then a note of that was made in the comment field. Once all of the files had been aggregated into one file, an overview file was created that contained a link to the original "high bill" call file, the actual player name in KITT, the account number, the utility company associated with the account, the date of the call, and any relevant comments. An attempt was made to create a link to each KITT account for each player, but it turns out this is not possible, according to one of the KITT administrators. This is to help prevent potential security issues.

From here the data needed to be cleaned up. Some of the customer service representatives reported some, but not all, of the information listed above. When this was the case, the default values for hours per day, hours per month, and kWh per month were used, and Watts was calculated as kWh per month divided by hours per month and multiplied by 1000. There were a number of appliance categories that were technically the same thing, but they were spelled differently or included irrelevant details, so they needed to be aligned. For example, "Air Conditioner" would be the same thing as "AC", and "Plug Load (electronics, etc. in standby mode)" would be the same thing as "Plug Load" and "plug load." These were all minor nuances but they were issues that needed to

be addressed. Google Refine[1] was used to rename the appliances so all of the accounts had common appliance names.

The various people answering the "high bill" calls would ask different questions. Moreover, all of them had a different proficiency with Excel, ranging from inputting data, to creating graphs and valuable output to provide to the customer. These discrepancies made it obvious that a better and easier starting point for all "high bill" calls needed to be determined. The beginning work was undertaken for a more user friendly worksheet that customer service could use when people called about a "high bill." This included creating push buttons, using macros that would eliminate any appliances that the customer said they did not have in their home, combining similar appliances, and eliminating some of the smaller ones. This way a "high bill" call could get to the most relevant information more quickly. To better accomplish this, meetings were set up with a few of the customer service representatives to obtain information on the process of "high bill" calls. The goal of these meetings was to find out if a certain age group tended to make these calls, and if there were certain appliances that tended to be the reason behind a "high bill" call. The information obtained was used to begin creating the new "high bill" call worksheet. This task was eventually completed by another EM&V (Evaluation, Measurement, and Verification) coworker.

APPLIANCE PREDICTION ANALYSIS

I explored various techniques used to predict which appliances an account might have, based on the entire account's electric usage. This was attempted using data from a group

[1] Google Refine (now Open Refine) is a power tool for working with messy data, cleaning it up, transforming it from one format into another, extending it with web services, and linking it to databases.

of high resolution homes. A high resolution or high performance home is a home that is highly energy efficient by using some of the most cutting edge appliances (most homes also had solar panels and other devices that generated their own electricity). Each of these homes had agreed to have the majority of their appliances monitored with an eMonitor. These monitors would read the electric usage of each appliance and the entire home's usage every minute and report it to SiteSage[2].

INITIAL ANALYSES WITH HIGH PERFORMANCE HOME DATA

Originally the only way to obtain the high resolution data from the high performance homes was to download two weeks of one-minute data from the SiteSage website for each home individually. This was done for three of the twenty four accounts, to obtain a six month period of data from each account. Once all of the data had been obtained in two week increments, each Excel worksheet was combined together. This data was explored using many different functions in order to begin to look at relationships between appliances and accounts. A function was created to expand the temperature readings that occurred once every hour to cover the entire hour so that there could be a temperature variable. The one-minute data was also aggregated into 15-minute data because 15-minute data is what VEIC is going to be receiving in the future. This made it possible to see if there was any significant difference in using one-minute versus 15-minute data. Another thing that needed to be done was to clean up the data and change the appliance names so they would line up better across accounts. Next, density plots, histograms, correlation plots, and general linear models using binary variables (appliance on or off)

[2] SiteSage is the company that obtains and stores all of the data from the eMonitors installed across Vermont.

for the data from the three accounts were generated. Clustering was attempted using both binary and numeric variables. Most of the knowledge regarding clustering came from "Finding Groups in Data: An Introduction to Cluster Analysis" by Kaufman and Rousseeuw. After looking at the data, a list was made of the major appliances that were believed would be the focus of the study. These included the hot water heater (DHW), dishwasher, dryer, range, range hood, heat pump (HP), microwave, refrigerator, ventilation system (Vent), washer, and the well pump (WP). Tree classifications were then considered for these major appliances using density and tree plots. These classifications were attempted using both the numeric values for the wattages of the appliances and a binary variable for on/off. Once the tree classifications were performed, it became apparent that the hour of the day might have an impact on the usage of various appliances. For example, an electric water heater might be used more during the morning and before bed rather than late at night and in the middle of the day. This would result in possible correlations with the time of day and the usage for some routine appliances. To look at this possibility, a function was created that broke the data for an account into hourly data, and then I reran the tests mentioned above on this new data. Heat maps of the data were also created to look at some of the patterns between appliances, the time of day, and the time of year. I had discussed random forests with some of the other EM&V employees, and it was thought that they might be applicable to the data here, so they were attempted. As a result of a presentation of the histograms, heat maps, and correlation plots of the high resolution data, it became necessary to look for a better way to obtain the high performance home data from SiteSage. It took a while to get all of the required

paperwork completed, but eventually an API was set up in order to allow the data to be downloaded from SiteSage using an R script.

API DOWNLOAD

This section deals with the ability to download the data for the high performance homes from the SiteSage website and some problems that arose during the process. VEIC created and signed a contract with SiteSage allowing the data for the high performance homes to be acquired from SiteSage using an API[3], rather than downloading them in two week increments. The first thing that I needed to do was to read the manual that SiteSage provided on how to call the API to get the desired data. This information was cross referenced with R to figure out which, if any, packages could be used to do this. From this research it was found that the RCurl package was the most common package allowing an API to run through R. Unfortunately, there was a problem and some of the functions did not work with the SiteSage website so, another way to access the data needed to be found. The httr packages were discovered, after much trial and error, which worked well when tested in R. Once it was known that it was going to be possible to obtain the desired data, and there was a way to do it, I created an Excel spreadsheet that contained all of the relevant information for each home, so it could be drawn from when retrieving data from the API. This information included the account name, the serial number on the monitor, and the date that the monitor started recording data. A function was created to read all of the data from the SiteSage website, take all of the desired information, and write that to a .csv file for future use. Due to the amount of data that

[3] An API is a software intermediary that makes it possible for application programs to interact with each other and share data. It is often an implementation of REST that exposes specific software functionality while protecting the rest of the application.

was being retrieved from the SiteSage website, they would only allow one day's worth of one-minute data for one account to be downloaded at a time. The biggest problem I found when running this script was that R would run out of memory, because it was dealing with millions of data points at a time. I made various endeavors to rectify this situation by adjusting the created function to retrieve the data from SiteSage. On the fourth attempt, I discovered the ability to append the data retrieved to a .csv file and then delete it from the system to free up some memory. Unfortunately, R does not have the ability to completely free up the used space unless you restart the program, so this only prolonged my ability to use R. I also found that the gc() function helped to clear space in the memory as well. The final function that was created first retrieved the security key using a username and password. Then it created a series of days to loop through so all of the desired data could be retrieved. The start date and the end date were input into the function so that, if and when R ran out of memory, the function could be restarted where it left off by changing the start date. Next, the function made sure that the security key worked, and if it did not, then it asked for a new one. Once there was a functioning security key, it recovered the information for the first day in raw text format. Then the function parsed through this raw data to create a data frame containing the appliance names, the time of the reading, and the usage of each appliance monitored. Next, it took this data frame and appended it to a .csv file chosen by the user. Finally, it deleted anything that was irrelevant and repeated this process for the next day until it either reached the end of the series of days or R ran out of memory. Within this function a progress bar was also created so the user could see how much longer it was going to take

for R to complete the process, or how far it got before it ran out of memory. This function was run on all 24 accounts from which data could be retrieved. There was a considerable amount of data, and it took four computers over a week to download everything. The functions had to be restarted throughout the week as R ran out of memory. After all of the data had been read from the SiteSage website, it was necessary to add NA values for missing readings and daylight savings time as the API simply skipped them. As a result, all of the data sets would have the same number of readings in a day. This script has since been updated and was run to obtain updated data sets on the original accounts as well as data on the twenty new accounts that have been activated through SiteSage.

APPLIANCE PREDICTION ANALYSIS CONTINUED

Once all of the data had been retrieved through the API, some analyses were performed on it. This started by reading in all of the data and then reducing each account down to the previous six months (January to August) so everything would be on the same time frame. Some accounts ended before or during this period so they were not used. This left 15 accounts; however, 3 of them had a large chunk of data missing during the middle of the time period, which was not noticed until later. The names of the appliances were changed to a more conventional naming scheme so all of the accounts had matching variable names. The appliances that were recorded twice or were extremely similar (for example: range, oven, and cooktop were considered the same appliance) were combined into one variable (this was done according to the information from VEIC's liaison with SiteSage). Appliances were recorded twice if it took more than one monitor to obtain the

total usage for an appliance (this happened for some of the bigger appliances and the main power). Devices that contained negative values or were not used during the six month period were eliminated, since they should not be accounted for in the main power (the negative values arose from appliances that generated electricity like a solar panel). The negative usage was added back into the main power variable to make up for this elimination. Again, it was decided that it would be a good idea to aggregate the data for each account up to 15 minute data, because it would be more like the data that VEIC would get in the future. This way a comparison could be made as to how much better, if at all, the one-minute resolution data would be compared to the 15 minute data. Correlation plots were then created for appliances within accounts and across accounts for both the one-minute and the 15 minute data. The correlations within accounts were the correlations between each appliance in a single account with the main power of that account, while the correlations across accounts were the correlations for a single appliance and the same appliance from different accounts. These correlations were plotted (Figure 6) to see which, if any, appliances were highly correlated with other appliances or with the same appliance across accounts. It became apparent that it might be a good idea to look at all of the data in each hourly time period, so the one-minute resolution data was broken down by hour, and the correlation plots were repeated to see if anything changed. Heat maps were created next to see patterns for different appliances by the time of day and by the day of the year. Different variations of the heat maps were created to explore how different accounts would look with and without certain appliances, as well. These heat maps were used on both the one-minute and the 15

minute data, but there was no significant difference in the two. I also looked into clustering the data into two or three groups based on the total usage of the home. This worked well for some of the accounts but not for all of them. Tree plots were used to split the data into different categories, but again this worked well for some accounts but not all of them. The last graphical analysis was to plot (Figure 7) each account's average DHW usage per day on an overlay plot, in order to show that some accounts had a very high usage, while some did not use the DHW much at all.

PREDICTING COMMERCIAL ACCOUNT ELECTRIC USAGE

This section deals with the data for the commercial accounts in the state of Vermont. The first part explains what was done to the data to be able to predict electric usage. The second part deals with the analysis of the various prediction methods used.

ORGANIZING THE DATA AND APPLYING THE MODELS

The commercial account electric usage prediction project involved looking at data for over 38,000 commercial accounts across Vermont, as well as the average daily temperature in Burlington. The commercial account data was obtained from an EM&V intern, and some adjustments were made to make it easier to work with (changing the names of variables, etc.). The daily temperature for Burlington was found from the University of Dayton website. Using 55 and 65 as the heating degree day and cooling degree day base temperatures respectively, the heating and cooling degree days for everyday, based on the average temperature in Burlington, were calculated. Then all the data for the accounts that had fewer than 36 read dates was removed as these accounts did not provide enough information for the models to perform adequately. Next, the number

of heating and cooling degree days between each reading period was calculated for all of the commercial accounts and added as variables to the data. Once all of the necessary data had been obtained, it was time to begin applying the models to determine which one was the best model for predicting future usage, based on past usage and temperature. I attempted a number of techniques using R, but there did not appear to be a good way to apply each model to each account individually without errors. Eventually, I stumbled upon the split-apply method, in which you would turn each account into a list element and then apply each model across the entire list, one element at a time. This method was tested on a couple of small data sets where it worked perfectly, so the split-apply method was used on the commercial account data. A number of models were run on the data. The six models used were: the Normal, the Poisson, the Log Normal, the Robust Normal, the Robust Poisson, and the AutoRegressive Integrated Moving Average (ARIMA) models. The lapply() function in R was used to run each model over the list of accounts. Each account was broken up into a testing and a training data set. The testing set contained the last twelve readings (these were roughly the readings for the last year of data), and the training data set contained the rest of the data. Due to some calculation issues (e.g., one of the accounts had zero variability which caused an issue with the ARIMA model) with some of the models, some of the data needed to be slightly adjusted so the model could be performed across all of the accounts without running into an error. Each model was tested with the training set for each account and then was used to predict the values of the testing set. The ARIMA model was run using the auto.arima() function from the forecast package in R. This function would look at the data and determine what

it believed to be the best ARIMA model based on some initial criterion provided by the user (e.g., maximum number of differences to look at). A few error statistics were calculated for each model to analyze the differences in performance across models. These included the mean absolute error (MAE), the minimum error, the maximum error, the total absolute differences, and a "results" statistic that was equal to the sum of the predicted values minus the sum of the actual values, all divided by the sum of the actual values. While each of these statistics was considered, only the "results" statistic is presented. The choice of the "results" statistic arose because the overall results for each of the measured statistics were nearly identical. All of the information from each model was returned to a new list where all of the relevant data was extracted and written to a .csv file to be used in further analysis. This included the Customer Account numbers, the average kWh for the training and testing data sets, the maximum and minimum kWh from the testing data set, the mean absolute error, the minimum and maximum error, the "results" statistic, the absolute differences, the coefficients from the models, and the aic, and arima components (when applicable). From here a new script was created to look at the results.

COMPARING THE MODELS

This script read in all of the data from the .csv files and began by plotting the data in 10% bins and looking at various statistics to find patterns. Functions were created that would calculate confidence intervals and bootstrapping intervals for the data. For each model a subset was created with the top 80% of businesses, based on the average kWh from the training data sets. The confidence interval and bootstrapping interval functions were run

for sample sizes of 5, 10, 15, 25, 50, and 100 on the “results” statistic. These intervals were plotted to see how the different models performed. The correlations between the heating and cooling degree day coefficients from the models were investigated but there was nothing statistically significant about them. Bootstrap intervals for the same sample sizes for the mean absolute error were also created, only to find similar results. For the ARIMA model data, the number of unique ARIMA models was also determined.

RESULTS

DATA AGGREGATION FOR "HIGH BILL" CALLS

The "high bill" data was cleaned, organized, and aggregated into one large spreadsheet with all of the pertinent information from each call. Figure 1 contains a sample of two files before the aggregation process and Figure 2 shows what these would look like after aggregation.

FIGURE 1.1 PRE-AGGREGATION EXAMPLE 1

1	Electric Rate:	0.15		hrs/month:	720	0.666666667			Typical
2									Usage
3	**Appliance**	**Watts**	**Qty**	**Hrs/Day**	**Hrs/month**	**kWh/month**	**Cost/month**	**Hrs/Day**	**Hrs/month**
4	Air Conditioner - Central System	3,000	0	4.2	125	0	$ -	4.17	125
5	Air Conditioner - Window Unit 8,000 BTU *	900	0	3.3	100	0	$ -	3.33	100
6	Air Purifier-Electrostatic	50	0	24.3	730	0	$ -	24.33	730
7	Air Purifier-Standard	50	0	24.3	730	0	$ -	24.33	730
8	Answering Machine (not combined with cordless phone	7	0	24.3	730	0	$ -	24.33	730
9	Aquarium w/ Heater, Light, Filter	95	0	12.0	360	0	$ -	12.00	360
10	Clock Radio	1	0	24.3	730	0	$ -	24.33	730
11	Clothes Dryer, Electric (6 loads per week)	3,750	1	0.7	20	75	$ 11.25	0.67	20
12	Clothes Dryer, Gas[1] (6 loads per week)	400	0	0.8	23	0	$ -	0.77	23
13	Clothes Washer[2] (7 loads per week) *	300	1	1.0	30	9	$ 1.35	1.00	30
14	Coffeemaker (10 pots per week)	150	1	1.0	30	5	$ 0.68	1.00	30
15	Computer with Monitor	125	1	2.0	60	8	$ 1.13	2.00	60
16	Crock Pot	200	0	0.1	3	0	$ -	0.10	3
17	Dehumidifier (damp basement) *	800	0	12.0	360	0	$ -	12.00	360
18	Dishwasher - Air Dry (4 loads per week) *	500	0	0.5	16	0	$ -	0.53	16
19	Dishwasher - Heat Dry (4 loads per week) *	800	0	0.5	16	0	$ -	0.53	16
20	DVD Player or VCR	23	2	0.5	16	1	$ 0.11	0.53	16
21	Electric Blanket (queen size)	35	0	8.0	240	0	$ -	8.00	240
22	Engine Block Heater	750	0	6.0	180	0	$ -	6.00	180
23	Fan - box or floor stand	150	0	2.4	71	0	$ -	2.37	71
24	Fan - ceiling (without lights) *	80	0	5.0	150	0	$ -	5.00	150
25	Freezer Chest, 17 CF, manual defrost, new *	49	0	24.3	730	0	$ -	24.33	730
26	Freezer Chest, 18 CF, manual defrost, 10 years old	70	0	24.3	730	0	$ -	24.33	730
27	Freezer Chest, 18 CF, manual defrost, 20 years old	102	0	24.3	730	0	$ -	24.33	730
28	Freezer Upright, 17 CF, auto defrost, 10 years old	124	0	24.3	730	0	$ -	24.33	730
29	Freezer Upright, 17 CF, auto defrost, 20 years old	153	0	24.3	730	0	$ -	24.33	730
30	Freezer Upright, 17 CF, auto defrost, new *	78	1	24.3	730	57	$ 8.56	24.33	730
31	Freezer Upright, 17 CF, manual defrost, 10 years old	69	0	24.3	730	0	$ -	24.33	730
32	Freezer Upright, 17 CF, manual defrost, 20 years old	105	0	24.3	730	0	$ -	24.33	730
33	Freezer Upright, 17 CF, manual defrost, new *	55	0	24.3	730	0	$ -	24.33	730
34	Furnace Fan	856	1	6.0	180	154	$ 23.11	5.93	178
35	Garage Door Opener (1/2 HP motor)	375	0	0.0	1	0	$ -	0.03	1
36	Garbage Disposal (1/2 HP motor)	375	0	0.0	0	0	$ -	0.00	0
37	Hair Dryer (10 minutes per day)	1,250	1	0.2	5	6	$ 0.94	0.17	5
38	Heat Lamp	250	0	0.1	2	0	$ -	0.05	2
39	Heat Recovery Ventilator (HRV)	125	0	10.0	300	0	$ -	10.00	300

Figure 1.1 is a good example of a file that contains a number of appliances with a quantity of zero as well as appliances that are nearly identical, (e.g., the three upright freezers).

FIGURE 1.2 PRE-AGGREGATION EXAMPLE 2

3	Appliance	Qty	Hrs/month	kWh/month
4	Clothes Dryer, Gas (6 loads per week)	1	30	12
5	Clothes Washer (7 loads per week)	1	12	4
6	Computer with Monitor	1	120	15
7	Dehumidifier (damp basement) *	1	360	288
8	Dishwasher - Heat Dry (4 loads per week) *	1	16	13
9	DVD Player or VCR	1	16	0
10	Fan - ceiling (without lights) *	1	150	12
11	Freezer Chest, 18 CF, manual defrost, 20 years old	1	730	75
12	Freezer Upright, 17 CF, manual defrost, new *	1	730	40
13	Heating System - hot water circulator (3 zones)	1	180	49
14	Lighting - 20W compact fluorescent bulb (75W equivale	5	100	7
15	Oven or Broiler	1	9	24
16	Plug Load (electronics, etc. in standby mode)	1	730	37
17	Refrigerator - 22 CF, side-by-side, 20 years old	2	730	270
18	Satellite/Cable Receiver Box *	1	730	18
19	Telephone - cordless with digital answering machine	1	730	5
20	Television - 32" to 60" LCD flat screen *	3	150	68
21	Well Pump (~40 minutes a day)	1	18	12
22				947
23				
24				
25				
26				
27	Appliance	Qty	Hrs/month	kWh/month
28	Well Pump (5 hours a day)	1	150	103

Figure 1.2 is more organized but it contains extra information for some appliances (e.g., the television), and there are two separate well pumps that were running for different amounts of time per day.

FIGURE 2.1 AGGREGATION EXAMPLE 1

1	PlayerName	Account	Appliance	Qty	Hrs/Day	Hrs/month	Watts	kWh/month	Date of Call	Comment	Hrs/Day	Hrs/month
143	RD	90-****	Clothes Dryer, Electric	1	0.7	20	3750	75	2/1/2013		0.7	20
145	RD	90-****	Clothes Washer	1	1	30	300	9	2/1/2013		1	30
146	RD	90-****	Coffeemaker	1	1	30	150	5	2/1/2013		1	30
147	RD	90-****	Computer with Monitor	1	2	60	125	8	2/1/2013		2	60
152	RD	90-****	DVD Player or VCR	2	0.5	16	23	1	2/1/2013		0.5	16
162	RD	90-****	Freezer Upright	1	24.3	730	78	57	2/1/2013		24.3	730
166	RD	90-****	Furnace Fan	1	6	180	856	154	2/1/2013		5.9	178
169	RD	90-****	Hair Dryer	1	0.2	5	1250	6	2/1/2013		0.2	5
174	RD	90-****	Heater - portable	1	24	168	1100	185	2/1/2013		8	240
183	RD	90-****	Lighting	10	3.3	30	13	4	2/1/2013		3.3	100
184	RD	90-****	Lighting	2	3.3	30	60	4	2/1/2013		3.3	100
188	RD	90-****	Microwave Oven	1	0.3	8	1427	11	2/1/2013		0.3	8
191	RD	90-****	Oven or Broiler	1	0.3	9	2660	24	2/1/2013		0.3	8
194	RD	90-****	Plug Load	1	24.3	730	50	37	2/1/2013		24.3	730
201	RD	90-****	Refrigerator	1	24.3	730	55	41	2/1/2013		24.3	730
205	RD	90-****	Satellite/Cable Receiver Box	1	24.3	730	25	18	2/1/2013		24.3	730
209	RD	90-****	Telephone	1	24.3	730	7	5	2/1/2013		24.3	730
210	RD	90-****	Television	1	5	150	120	18	2/1/2013		5	150
212	RD	90-****	Television	1	5	150	150	23	2/1/2013		5	150
214	RD	90-****	Toaster or Toaster Oven	1	0.1	3	1200	4	2/1/2013		0.1	3
217	RD	90-****	Video Game Box	1	0.8	24	175	4	2/1/2013		0.8	24
221	RD	90-****	Well Pump	1	1.2	36	686	25	2/1/2013		0.6	17

FIGURE 2.2 AGGREGATION EXAMPLE 2

1	PlayerName	Account	Appliance	Qty	Hrs/Day	Hrs/month	Watts	kWh/month	Date of Call
582	AJ	408****	Clothes Dryer, Gas	1	1	30		12	10/24/2013
583	AJ	408****	Clothes Washer	1	0.4	12		4	10/24/2013
584	AJ	408****	Computer with Monitor	1	4	120		15	10/24/2013
585	AJ	408****	Dehumidifier	1	12	360		288	10/24/2013
586	AJ	408****	Dishwasher - Heat Dry	1	0.5	16		13	10/24/2013
587	AJ	408****	DVD Player or VCR	1	0.5	16		0	10/24/2013
588	AJ	408****	Ceiling Fan	1	5	150		12	10/24/2013
589	AJ	408****	Freezer Chest	1	24.3	730		75	10/24/2013
590	AJ	408****	Freezer Upright	1	24.3	730		40	10/24/2013
591	AJ	408****	Heating System	1	6	180		49	10/24/2013
592	AJ	408****	Lighting	5	3.3	100		7	10/24/2013
593	AJ	408****	Oven or Broiler	1	0.3	9		24	10/24/2013
594	AJ	408****	Plug Load	1	24.3	730		37	10/24/2013
595	AJ	408****	Refrigerator	2	24.3	730		270	10/24/2013
596	AJ	408****	Satellite/Cable Receiver Box	1	24.3	730		18	10/24/2013
597	AJ	408****	Telephone	1	24.3	730		5	10/24/2013
598	AJ	408****	Television	3	5	150		68	10/24/2013
599	AJ	408****	Well Pump	1	0.6	18		12	10/24/2013
600	AJ	408****	Well Pump	1	5	150		103	10/24/2013

Figure 2.1 and 2.2 show what the two examples would look like after aggregation.

This spreadsheet can be used to find patterns between appliances and “high bill” calls. This information may lead to increased knowledge about which types of people, which types of appliances, and what times of the year are the most likely to result in making a “high bill” call.

APPLIANCE PREDICTION ANALYSIS

Figure 3 contains an example of the histograms and correlation plots for one of the high performance home accounts.

FIGURE 3.1 HISTOGRAMS

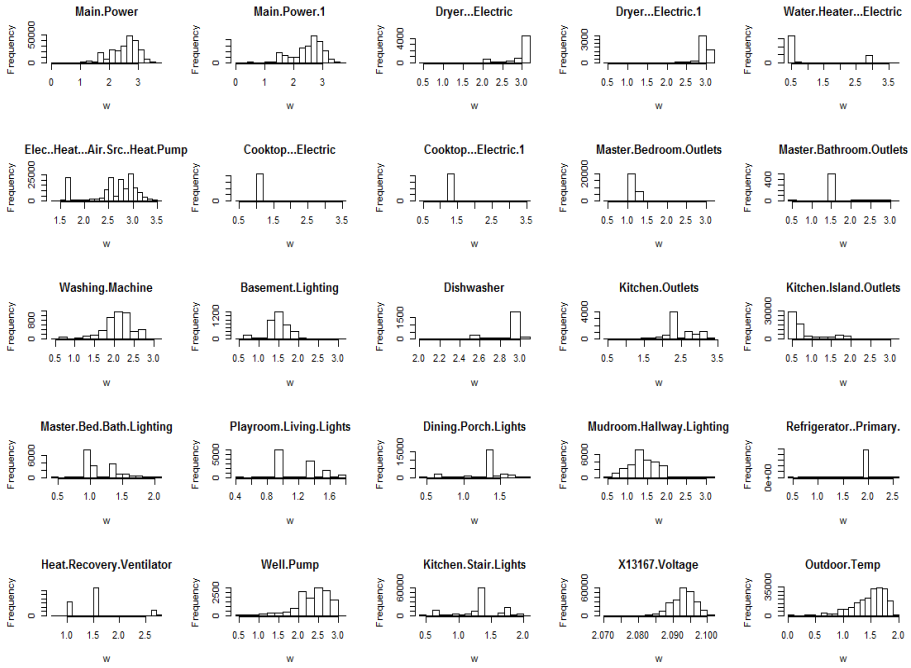


Figure 3.1 shows the histograms for the logged wattages for each of the monitored appliances as well as for the outdoor temperature. The zero wattage values were eliminated because they dominated the majority of the appliances and obscured the rest of the outcomes. As you can see, there are some appliances that are listed more than once. This was an issue that had not been dealt with at this point, but it was fixed once delving deeper into the high resolution home data. According to VEIC's liaison with SiteSage, the solution was to add the measurements together, as these appliances required more than one monitor to measure the entire appliance output. From this figure, we can see which appliances have varying wattage outputs when they are on, and which appliances use the same amount of power every time. For example, the cooktop tends to use the same amount of power every time it is on, while the heat pump varies a great deal.

FIGURE 3.2 APPLIANCE CORRELATIONS

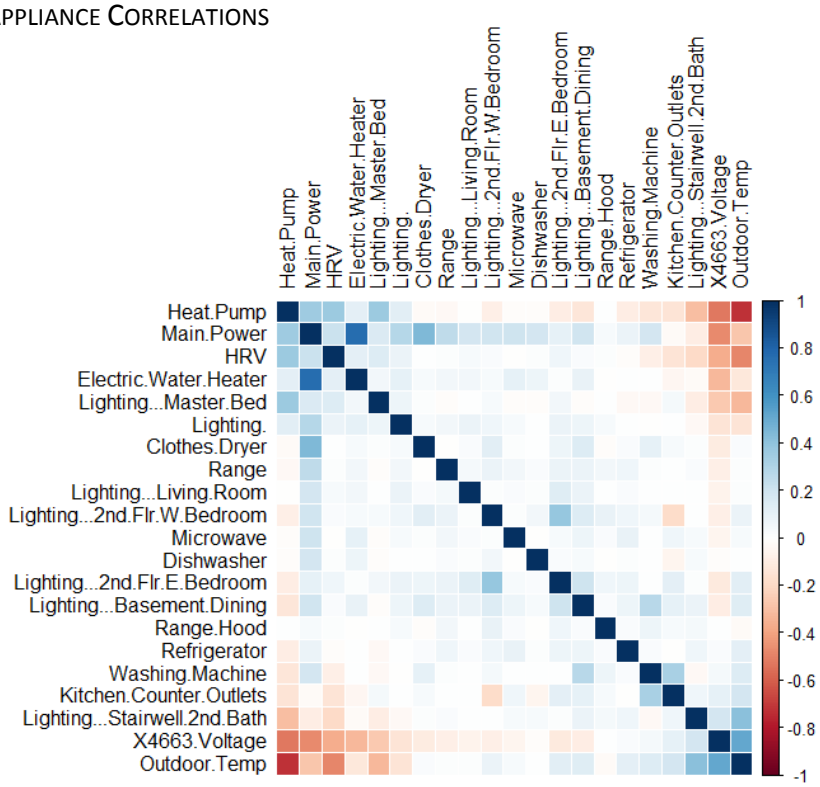
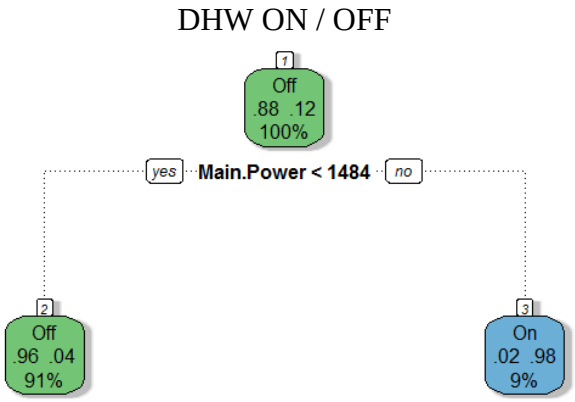


Figure 3.2 shows the correlations between appliances for one account throughout the six month period that the data was taken from. The dark blue boxes represent strongly positive correlations while the dark red boxes represent strongly negative correlations. From this figure we are able to see that the heat pump, water heater, HRV, range, microwave, dishwasher, and washing machine have positive correlations with the main power (with the water heater having the highest overall correlation). Figure 4 displays an example of one of the tree plots for a high performance home.

FIGURE 4 TREE PLOT



This basic tree plot was able to fairly accurately predict when the DHW was on or off, based solely from the main power of the entire house. This was not quite the goal of the project (predict whether an account actually has a DHW), but it was a good start and it generated ideas about ways to expand these results to better accomplish the goal. Clustering the data did not accomplish much, due to a lack of understanding the process enough to be able to include the results, and not fully understanding random forests caused problems getting any useful information from those results as well.

FIGURE 5 HEAT MAP

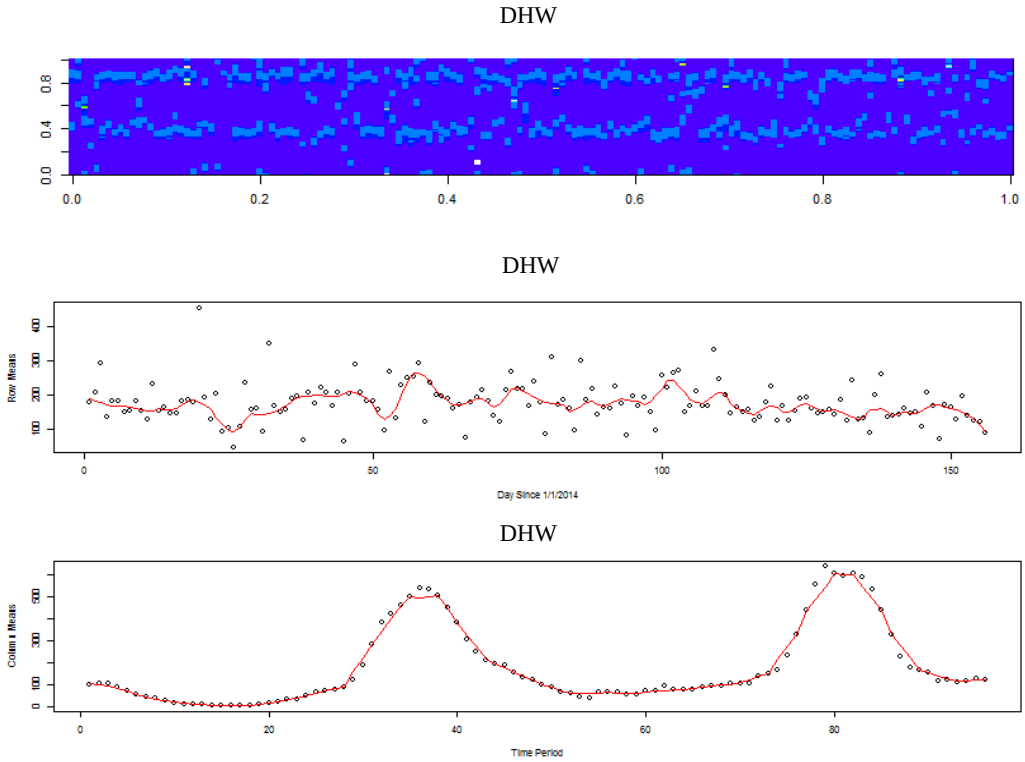


Figure 5 displays the heat map for the hot water heater in one of the accounts. It also includes the average usage across the six month period by day and by time of day in 15 minute increments with a lowess smooth of both plots in red. As can be seen in the bottom graph, the DHW was used mostly between 7:30am and 12:30pm and again between 5:30pm and 10:30pm. These heat maps helped to see when appliances were used for each account and to identify the most likely time that an appliance was being used. For example, the hot water heater (DHW) was normally used in the morning and then again at night in some accounts. This may be due to residents taking showers in the morning and the evening in some households, while other households only shower in the morning.

FIGURE 6 CORRELATIONS

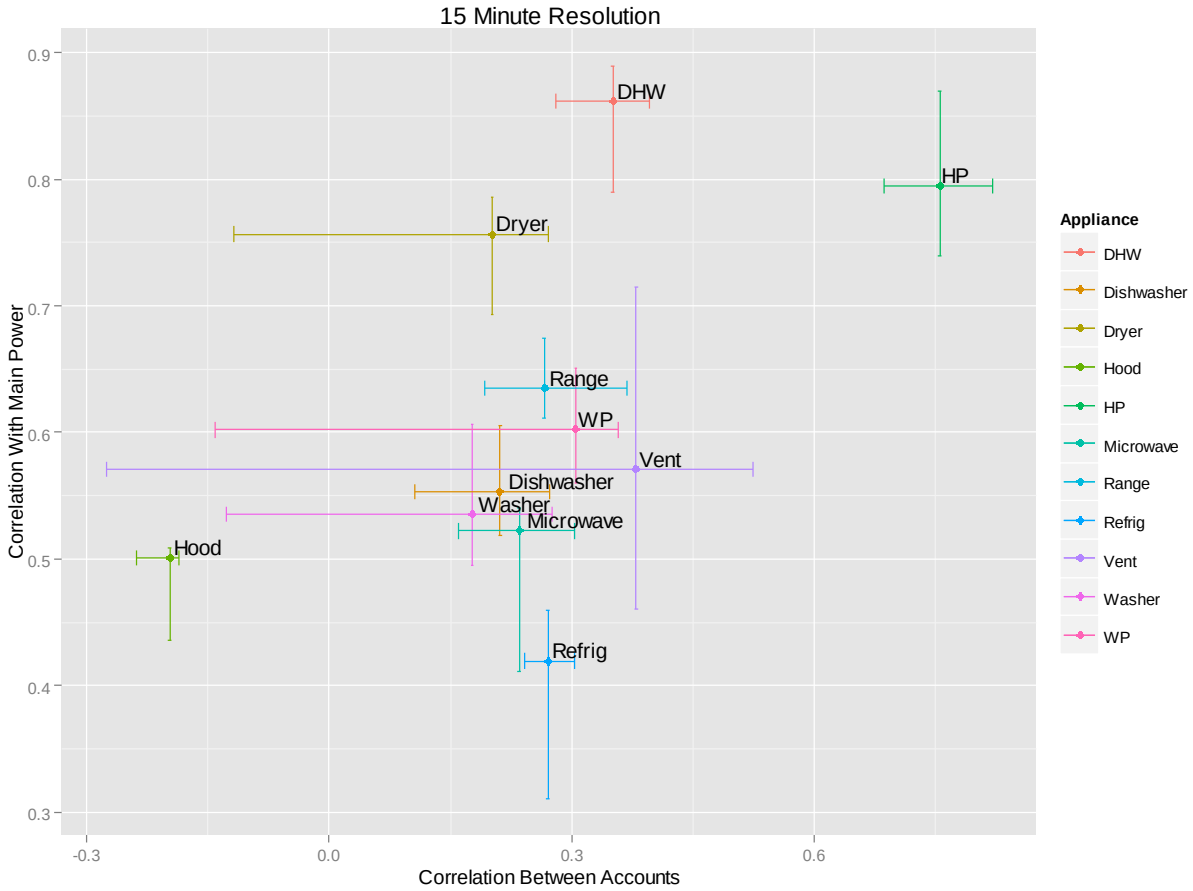
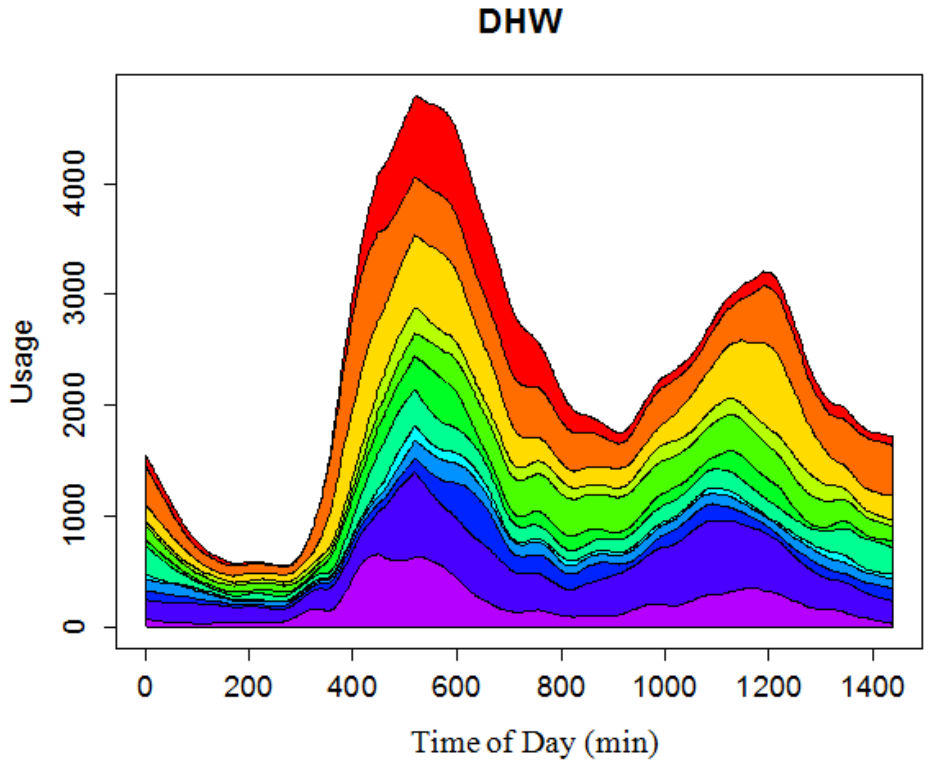


Figure 6 displays the cube root of the average correlations for each appliance with the main power from its account on the y-axis. The cube root of the average correlations for each appliance with the same appliance in other accounts is displayed on the x-axis. The cube root was chosen because the values of the correlations between accounts were all relatively small so this was easier to display. The lines extending from each point represent the interquartile range for each appliance's cube rooted correlations. Some of the appliances had high correlations within their accounts but only the heat pump had a higher correlation across accounts. It was believed that the appliances do not seem to correlate across accounts because different people might be using the same appliances at

different times of the day (different people have different habits). This helped to explain why the heat pump was highly correlated between accounts, as it is temperature dependent, and all of the homes were exposed to roughly the same weather patterns. The results had minimal change with the correlations within accounts when broken down on an hourly basis. However, some hours had higher correlations across accounts when looking at each hour individually.

FIGURE 7 DHW OVERLAY PLOT



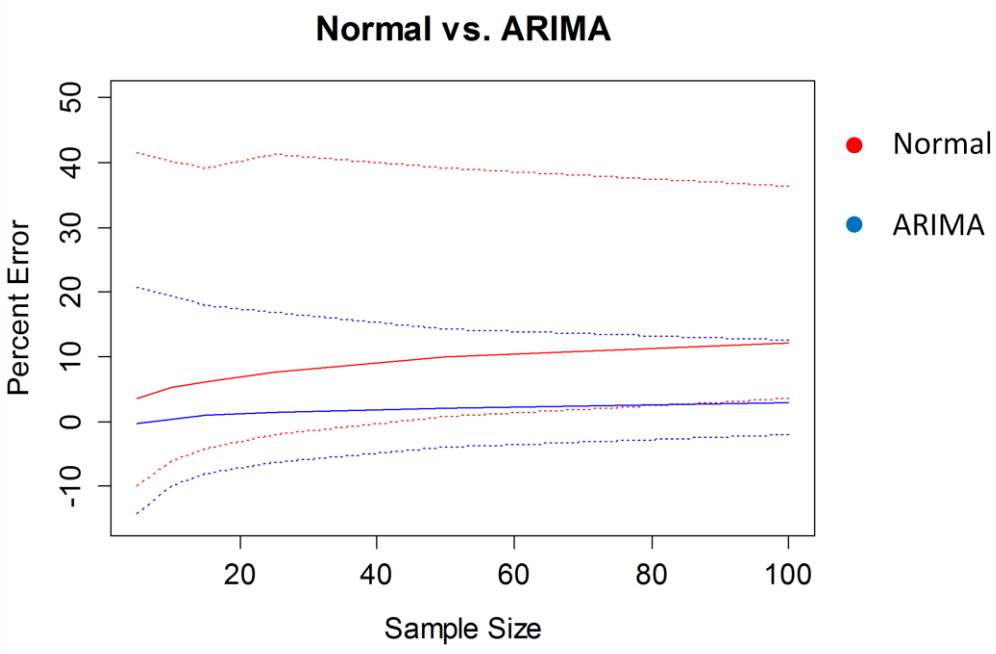
Each color from Figure 7 represents the average usage of the DHW for a different account throughout the day. This plot showed that in all of the accounts that had a DHW, it was in a low power state or off in the early morning hours, and on average it was using the most power between the 8:00am-10:00am and the 6:00pm-9:00pm time intervals.

Although the goal of the appliance prediction analysis was to be able to predict appliances based on total usage, no method was discovered that made accurate predictions.

PREDICTING COMMERCIAL ACCOUNT ELECTRIC USAGE

I discovered that 24131 out of the 38353 accounts had used some sort of ARIMA component, as opposed to the regular linear regression model (this is the default if auto.arima() does not find a trend or seasonality component). It appeared as though the ARIMA model performed the best compared to all of the other models, so the ARIMA bootstrapping results and the Normal GLM model bootstrapping results were plotted on the same graph to demonstrate how much better the ARIMA really was.

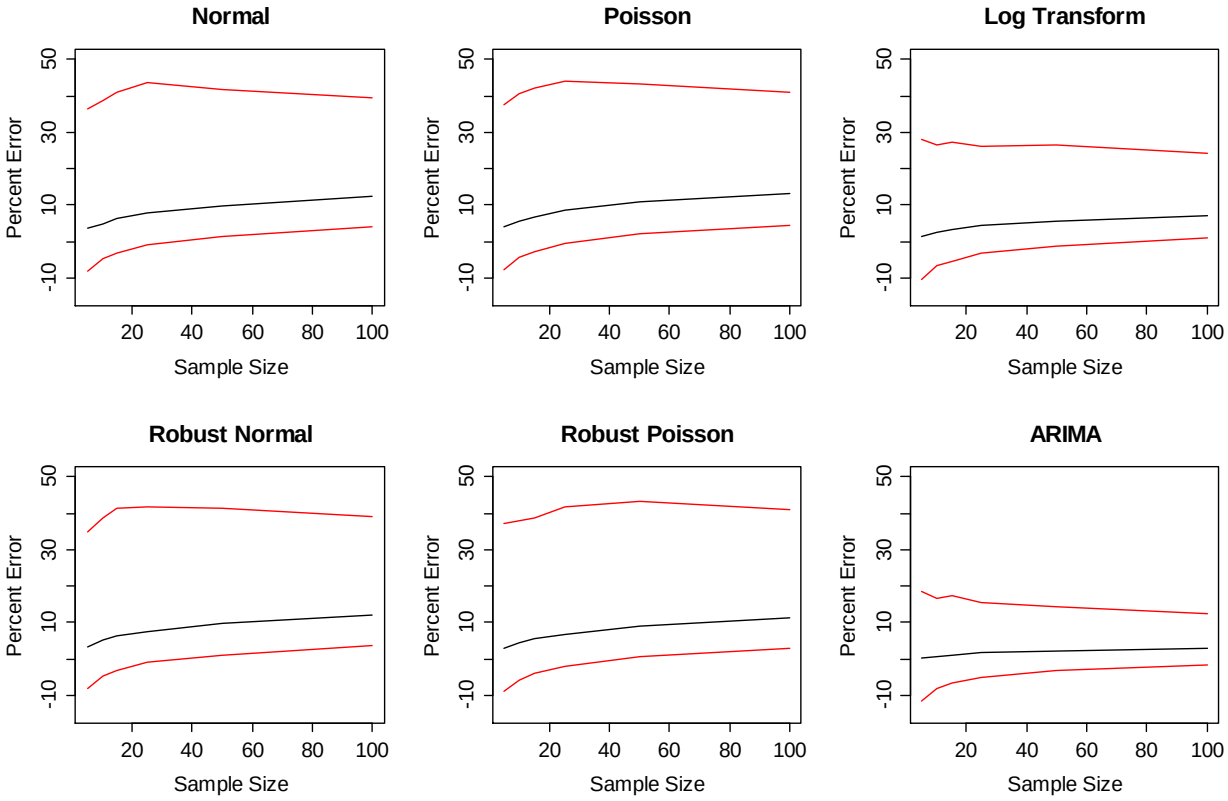
FIGURE 8 NORMAL VS. ARIMA MODEL



As can be seen from Figure 8, the ARIMA model outperformed the Normal model at all sample sizes and the interval estimates were much narrower, indicating a smaller

variance in the ARIMA model errors. The percent error was calculated using the "results" statistic. The AICs from the ARIMA model and the Normal GLM model were also compared, and it was found that 24754 of the AICs were better for the ARIMA while only 132 were better for the Normal GLM. This is another indication that the ARIMA model outperformed the Normal model when predicting future usage for the commercial accounts in Vermont. To illustrate how each model performed compared to the others, the bootstrapping plots of all six models were put together using the same axes for each model. This is displayed in Figure 9.

FIGURE 9 MODEL COMPARISON



For each model, the black line is the actual estimated percent error while the red lines represent the 80% bootstrapping interval for the percent error. As can be seen, the

ARIMA model contains the narrowest interval and remains the closest to zero for all sample sizes compared to all of the other models.

The go to method for prediction of future electric usage has been to use the Normal model on the data and predict usage based on the results. This paper found that, for the commercial accounts across the state of Vermont, the ARIMA model is more accurate in its predictions than the Normal model. Therefore, VEIC should begin to use the ARIMA model when they consult and engage in projects with businesses around the state.

CONCLUSION/FUTURE DIRECTION

The "high bill" call spreadsheet can be used in the future to examine relationships between different "high bill" calls. Smart meter data for each of these "high bill" calls can be obtained and then a baseline could be created, or found, for accounts that make "high bill" calls. This can then be used in conjunction with the appliance prediction analysis to attempt to predict which accounts are going to make "high bill" calls based on the appliances that are predicted for those accounts.

A major dilemma with the appliance prediction project was that different people had different routines and thus, even if an account possessed a DHW, the residents may not have used it for the same amount of time as another account, or they may have used the DHW multiple times throughout the day. There are a couple of techniques that should be looked into in order to overcome this issue. The first is dynamic time warping and the second is symbolic aggregate approximation. Each of these methods would essentially align the various usage habits for a given device across accounts. This would help to make different accounts as similar as possible and, therefore, make predictions easier to

accomplish. A method that might be used to actually predict whether an account possesses a certain appliance is logistic regression. This method could be used on the data available to create a predictive model for each appliance desired. The problem with this strategy is that there are not enough accounts to provide sufficient data to create an accurate model. However, there are now more than forty accounts being monitored by SiteSage, and as this number continues to grow it will become easier to create an accurate logistic regression model.

If we can accurately predict appliance presence then VEIC can participate in targeted marketing. For example, if a household has an electric water heater that is using a lot of energy, VEIC would be able to predict this and offer the household a rebate of some sort. This could be done without having to issue rebates to everyone in the hopes that they find the right person. These individual rebates could have values much larger than the widespread rebates since fewer people will be receiving them. Not only will this help to accomplish VEIC's mission of energy efficiency, but it will also cut costs associated with finding opportunities and implementing programs to achieve VEIC's goal.

The ARIMA model outperforms the current model utilized by VEIC so they should start using the ARIMA model instead when predicting future energy usage for commercial accounts in Vermont. This will lead to more accurate predictions of future electric usage and create better outcomes for the projects performed. By having an educated prediction as to which businesses are going to be utilizing a mass of energy in the following year, VEIC will be able to work with them to cut energy consumption, thus reducing costs, and increasing profits for these firms. If VEIC is successful in this endeavor, then other businesses may turn to them for consulting about future projects and other ventures. This will further advance VEIC's goal of increased energy efficiency.

References

Google Refine. [date unknown]. Google refine [Internet]. Google [cited in 2015 April 3]. Available from: https://code.google.com/p/google-refine/

Kaufman L, Rousseeuw P. 2005. Finding groups in data an introduction to cluster analysis. Hoboken (NJ): John Wiley & Sons Inc. 368 p.

Peng R. [updated 2014 May 31]. Computing for data analysis [Internet]. Youtube. [cited in 2015 April 3]. Available from: https://www.youtube.com/playlist?list=PL7Tw2kQ2edvpNEGrU0cGKwmdDRKcA6C4

Rouse M. [date unknown]. Essential guide to API management and application integration [Internet]. Essential Guide [cited in 2015 April 3]. Available from: http://searchcloudapplications.techtarget.com/definition/open-API

University of Dayton. [date unknown]. Average daily temperature archive [Internet]. University of Dayton web site. [cited in 2015 April 3]. Available from: http://academic.udayton.edu/kissock/http/Weather/citylistUS.htm