High Latitude House Electricity Consumption Analysis

Jingxiao Liu (jingxial)
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
412-805-2140
jingxial@andrew.cmu.edu

Junhua Wang (junhuaw1)
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
412-218-7131
Junhuaw1@andrew.cmu.edu

ABSTRACT
An accurate model to predict residential electricity consumption is a significant topic in the field of building energy management. These days, many researchers have developed several prediction models for both commercial buildings with huge population and residential buildings with small units of people. However, specific-site buildings provide particular consumption patterns caused by different environmental features. In this paper, we use a dataset collected from a high latitude house in Canada to establish a predictive model to analyze and describe the relationship between high latitude environmental features and electricity usage in one year.

To figure out the week periodic behaviors of one year's consumption data, cluster analysis is used. Regression tree model is suitable to be utilized here for modelling the consumption determined by climate features. Uni- and Multi-output regression tree models are tested by using K-Fold cross-validation method and fresh dataset to seek a better predictive model. Furthermore, to prevent overfitted model and calibrate the regression tree, different maximum depth is tested here. At the end, we conclude the total consumption of this high latitude house will be reduced when the global warming happens because of high consumption used by heater in winter. Then we separate the consumption of heater out of the whole house consumption to verify it.

CCS Concepts
• Computing methodologies→Regression trees; • Information systems→Data analytics; Clustering;

Keywords
Electricity, Climate Change, Cluster Analysis, Regression Tree

1. INTRODUCTION
Recently, electricity is consumed in almost every households around the world. Accurate model to analyze electricity consumption becomes an important topic in the field of building energy management. As is known that climate change, especially increasing temperature, has great influence on electricity consumption. For example, electricity industries have to satisfy the growth demands in summer when the number of extreme hot weather increases. Also, the electricity demands will decrease in high latitude area where will consume less electricity in winter for heating the house.

There have been several studies that have analyzed electricity data from smart meters across several homes or commercial buildings. Many of them have provided efficient models to analyze and predict energy usage across a large population [1]. However, the analysis of small unit, such as one household, is also important to provide necessary recommendations on energy management. Although the regular of electricity consumption is different for different household, there are some similar patterns of that in long period. For example, heater or air condition utilizes much more electricity with extreme cold or hot weather outside. Furthermore, in time domain, the consumption of household occurs time periodic every day or every week. Based on these discussion, questions we seek to answer are below:

1. How does electricity usage change daily and weekly in the same house? Does it indicate periodic regulation and relationship with some features?
2. How climate features effect electricity consumption, and which one has most influence on it?
3. Does climate change have influence on the electricity consumption of household? And how does it change in specific-site (Canada)?

Several approaches are used in this paper to depict electricity consumption dataset. Boxplots can describe the tendency of hourly electricity consumption in different weekdays; Cluster analysis is suitable to find the periodic difference of electricity usage; In order to classify climate features and seek the relationship between them with electricity consumption, regression tree is utilized.

The rest of the paper is organized as follows. In section 2, the dataset is briefly described. Section 3 presents amount of analyses applied to the dataset. Section 4 summarizes the analysis results and predicts consumption with increased temperature. Also, the predicted electricity consumption is compared with the real one. We conclude the paper and briefly discuss future work in section 5.

2. DATASET
In this Paper, the Almanac of Minutely Power dataset (AMPd), which contains hourly climate features, electricity, water, and natural gas consumption of a residential house in Canada from 2012 to 2014, is used to analyze and predict electric consumption in specific-site. AMPd data has been cleaned to provide for consistent and comparable accuracy results amongst different researchers and machine learning algorithms. It was collected from a house built in 1995 in the Greater Vancouver metropolitan area in British Columbia (Canada). The house is 80 m above sea level and the front of the house faces south. The house has one level above grade and a basement making up a total of 199 m^2. The main house has a family of three people: a male and a female adult in their late 30s and a daughter between the age of 5 and 6. The male adult is a full-time student at a local university, the female adult is self-employed, and the child attends full-time elementary school. A rental suite houses one male occupant in his early 20s with full-time employment [2].

This research is based on the minutely main house electricity, and hourly climate features (temperature, press, visibility, and weather) data. BC Hydro is the provincial utility that provides electricity to the house via a 240 V, 200 A service. Meter calibration was done

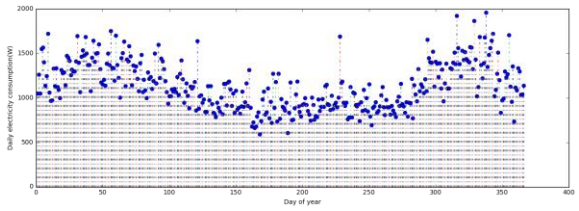

Figure 1. Daily electricity consumption (W) in one year.

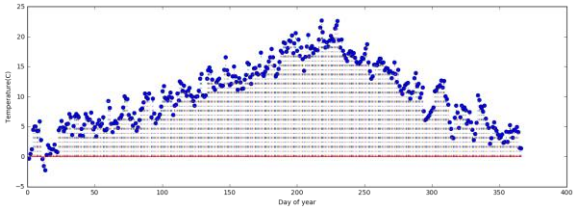

Figure 2. Daily temperature (℃) in one year

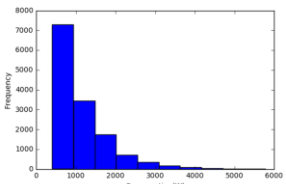

Figure 3. Histogram of the average hourly consumption.

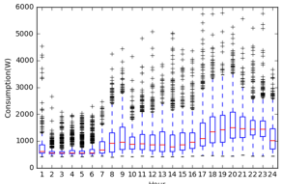

Figure 4: Boxplot of average hourly electricity consumption.

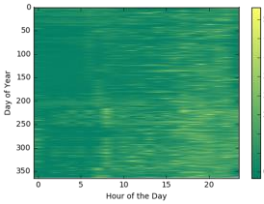

Figure 5. Heatmap showing average power consumption during an hour each day of the year of each day

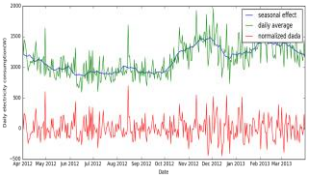

Figure 6. Curves of daily (green), seasonal (blue), and normalized (red) electricity consumptions

by the meter manufacture before shipping at the factory. Thus, in this paper, no further data cleaning works were processed.

3. Data Analysis

3.1 Periodic Phenomenon of Consumption

3.1.1 One Year Consumption and Temperature

Figure 1 is the stem plot which shows consumption (the y-axis) for each day in the first year of the dataset (the x-axis). That figure reveals that the trend of consumption is different in different seasons. In summer the consumption is lower but in other time of the year the consumption is higher. Figure 2 is the stem plot which shows temperature (the y-axis) for each day in the first year of the dataset (the x-axis). The figure shows that in summer the temperatures are between 15℃ and 20℃, which are comfortable for human, so there will be no need to use heater and AC. In other time in the year the temperatures are lower which means residents will use heater during this time.

3.1.2 Week and Daily Consumption

Figure 3 is the histogram showing the average hourly consumption on the entire dataset. It shows that the majority values of consumption are around 1000 W and more than 80% of consumption are less than 2000 W.

In subplots of 7 histogram plots which show the average hourly consumption in each day of a week. We find that there is no big difference between these 7 figures which means in weekdays and weekends the residents have similar electricity consumption habit.

Figure 4 is the box plot of average hourly electricity consumption for each hour of the day. From median value in box plot of each hour in a day, we find that the consumption from 0am to 7am has similar values which are very small. That is because residents are in sleep. From 8am to 9am, there comes a peak consumption. During this time residents wake up, and they would make breakfast. From 10am to 17pm have smaller values. Because residents are only at home in weekends. From 18pm to 23pm the consumption is higher, because residents are usually at home in this period, also when the temperature become low, it consumes more electricity for heating.

Furthermore, we find that in Saturdays and Sundays, from 8am to 24pm the values of electricity consumption are very similar because residents usually stay at home during daytime in weekends.

Figure 5 is the heat map showing average power consumption during an hour each day of the year (the y-axis) of each day in the first year in dataset (the x-axis), where dark colors indicate higher energy usage. The average hourly power of each day reveals the same morning, day, evening, and night trend as the box plots present. The figure also illustrates same seasonal trend of each day in a year.

3.2 K-means Cluster Analysis

K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. We use k-means cluster to distinguish the electricity consumption of weekdays and weekends. Our object is to find:

$$arg\min_{s} \sum_{i=1}^{k} \sum_{x \in s_i} ||x - u_i||^2$$

where $u_i$is the mean of points in s_i.

Assign each observation to the cluster whose mean yields the least within-cluster sum of squares (WCSS). Since the sum of squares is the squared Euclidean distance, this is intuitively the "nearest" mean. (Mathematically, this means partitioning the observations according to the Voronoi diagram generated by the means):

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \leq \left\| x_p - m_j^{(t)} \right\|^2 \ \forall j, 1 \leq j \leq k \right\},$$

where each $x_p$is assigned to exactly one $S^{(t)}$,even if it could be assigned to two or more of them.

BuildSys, Month 11–12, 2016, Pittsburgh, PA, USA.

DOI: http://dx.doi.org/10.1145/12345.67890

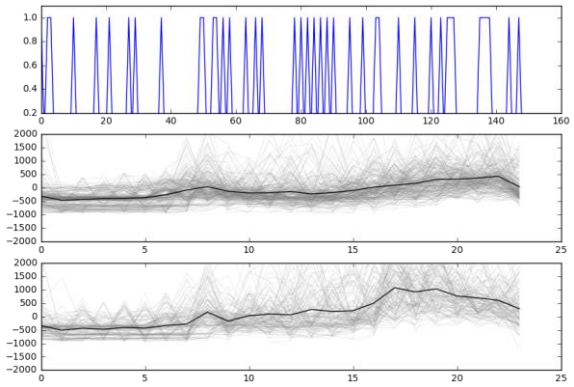

Figure 7. Load profile clusters (2 clusters applied) of buildings observed across the entire dataset.

Then the new means are calculated to be the centroids of the observations in the new clusters. [3]

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

In this paper, smooth window method is used to obtain the seasonal consumptions, which can help us to normalize the data. By subtracting seasonal data, the periodic regulation with small frequency can be shown and analyzed clearly. Figure 6 shows these three curves and the date time.

Then we take normalized data as input to do the k-means clustering with 2 clusters. The result is shown in Figure 7. We intend to find out the difference of electricity consumption values between weekdays and weekends. However, it seems that the consumptions of weekdays and weekends are randomly distributed in the first year. Not like commercial and social buildings which have predictive trend that weekdays would consume huge amount of energy and consumption of weekend is smaller, the residential house consumed electricity does not have big difference between weekdays and weekends.

3.3 Decision Tree and Calibration

We represent a dataset with N observations, where each input has n features and model has p outputs as:

$$x_i: [x_i^1, \dots, x_i^1]^T \in R^n,$$
$$y_i: [y_i^1, \dots, y_i^1]^T \in R^p,$$
$$i \in \{1,2, \dots, N\}.$$

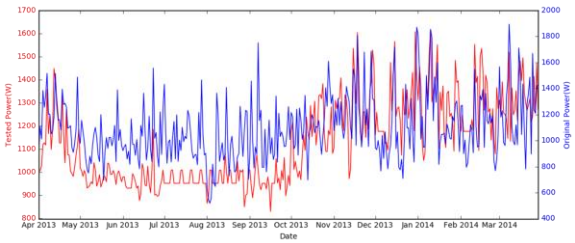

Figure 8. A comparison of real and predicted consumption in the second-year using regression tree developed from the first-year dataset.

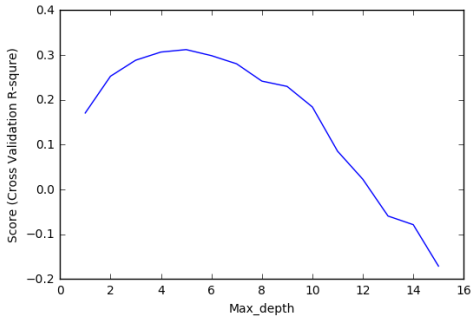

Figure 7. The curve of R-square score with different maximum regression depth.

And we get the optimal split at each node by minimizing the sum of mean square error in both the branches:

$$(x^k, t_k) = \arg\min \sum_{\{i|x_i \in R_L\}} (y_i - \overline{y_L})^2 + \sum_{\{i|x_i \in R_R\}} (y_i - \overline{y_R})^2$$

where $y_i \in R$ and $\overline{y_L}$ and $\overline{y_R}$ are the mean outputs of all the data points in R_L and R_R, respectively. The tree is grown in this fashion till the number of data points in the terminal nodes which contains observations less than the minimal number which is 2 in our case or reach the max depth of the tree. [4]

Initially, we take temperature, press, visibility, weather, hour, and weekday as features, and electricity power consumption as response values to create a regression tree model. Cross validation method with 10-fold is used to find the constraint of maximum depth. Figure 7 shows the change of R-square with different maximum tree depth. It can be noted that the regression tree model with cross validation method gets highest score when the maximum depth is 5.

4. Data Validation and Prediction

Section 3 had trained a regression tree for predicting electricity usage in one year. Then, we use the second-year data to test and validate the model. Also, we choose to use 5 maximum depths which can get highest score when applying to fresh dataset. R-square score is 0.275. The real second year consumption and predicted consumption are shown in Figure 8.

Furthermore, in this section, we increase the temperature with 3℃ as the new temperature feature intending to predict the consumption when temperature increases because of the global warming. Figure 9 shows the curves of original consumption and

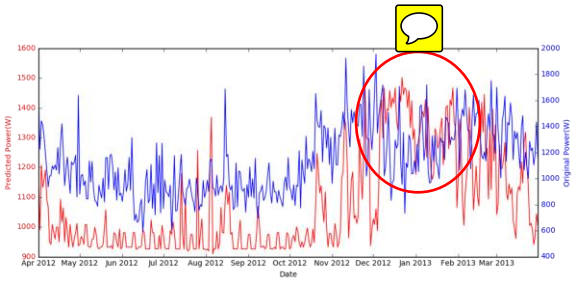

Figure 9. Curves of original consumption and predicted consumption with higher temperature in the first-year using regression tree developed from the first-year dataset.

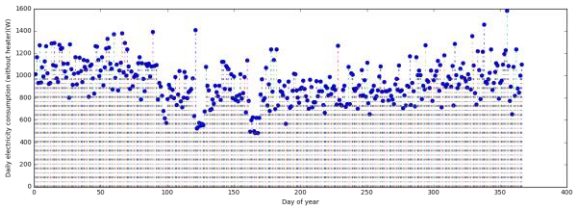

Figure 10. Daily non-heater consumption (W) in one year.

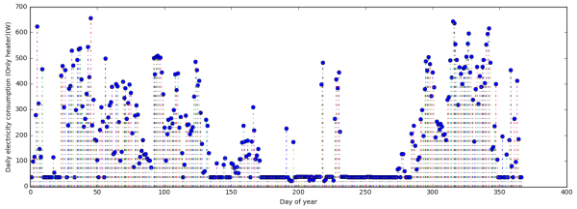

Figure 11. Daily heater consumption in one year

that with higher temperature in one year. The ratio of real consumption and predictive consumption with increasing temperature in every month and total are shown in Table1.

Table 1. Real consumption and predictive consumption with increasing temperature in every month and total.

Month	Real Consumption of the first year	Consumption with increasing Temperature	Ratio
1	973006	951559	1.02
2	825524	933624	0.88
3	827722	965652	0.86
4	735641	837054	0.88
5	711420	720260	0.99
6	684112	654236	1.05
7	721423	658215	1.10
8	736844	727283	1.01
9	684238	659815	1.04
10	744682	860090	0.87
11	805550	1000567	0.81
12	1002646	945615	1.06
Total	9452807	9913971	0.95

Although, those test scores are low, which is because that the consumption data in the first year is random, and it does not have strong relationship with the features we chose. It can be noted that with temperature increasing, electricity consumption in winter and the total consumption decrease. That can be caused by less usage of heater or air conditions, also, the highest temperature in summer is around 25 degree. Even by 3 degree increasing, people do not need to utilize the air condition in summer.

Finally, the whole house consumption is divided into heater consumption (Figure 10) and non-heater consumption (Figure 11). These figures show that there is no big difference across the first year when removing the heater consumption. Hence, heater consumption is the main factor that provides seasonal trend of consumption in the period of one year.

5. Conclusion and Future Work

In this paper, we conducted cluster analysis and regression tree model on electricity consumption collected from a house located in high latitude area where temperature is low. We identified and quantified the general trends and patterns observed in this individual house. The impact of weather, temperature visibility, press, and time on the electricity consumption are studied. Different from other area, the electricity usage of this house is higher in winter caused by huge heater consumption. We found out that this dataset has seasonal and daily trends, however, it does not present weekdays verse weekends pattern like that of commercial and social buildings. Further, we seek that the maximum depth has significant influence on regression tree model. A trained and validated model was used to predict consumption with increasing temperature. We concluded that the ratio of total consumption between real dataset and the predictive one is about 0.95, which means that the global warming makes the total consumption reduced in high latitude area. At the end, we removed the heater consumption from the whole house electricity consumption, and found out that it is the main factor that affects seasonal trend of consumption.

6. ACKNOWLEDGMENTS

Our thanks to Mario Bergés and Henning Lange for instructing us in the Data-Driven course.

7. REFERENCES

[1] Iyengar, S., Lee, S., Irwin, D., & Shenoy, P. (2016, November). Analyzing Energy Usage on a City-scale using Utility Smart Meters. In Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments (pp. 51-60). ACM.

[2] Makonin, S., Ellert, B., Bajic, I. V., and Popowich, F. 2016. Electricity, water, and natural gas consumption of a residential house in Canada from 2012 to 2014. Scientific Data, 3(160037):1–12.

[3] Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), 100-108.

[4] Jain, A., Behl, M., & Mangharam, R. (2016). Data Predictive Control for Peak Power Reduction.