CS506 Project Main Report

Topic: Mapping the Intensity of Energy Use Across Boston University
Group Members: Chengyu Deng, Xiaotong Niu, Qian Zhang
Due Date: Apr.12th, Thursday

Project Description

Well known for its extreme weathers, the city of Boston apparently neutralizes outdoor and indoor temperature with high energy consumption, either in winter by using a large amount of gas for heating or in summer by using a large amount of electricity for cooling. As a University located in the center of Boston, Boston University indeed follows the energy consumption pattern of the city of Boston. For this project, we are dedicated to analyze electricity and gas consumption data for buildings at BU, and then use Linear Regression approach to see how much weather contributes to energy consumptions and find factors that largely contribute to energy consumptions besides weather among the choices of water usage intensity, greenhouse gas emission intensity, building age and property type. In the end, we expect to have several factors that contribute to energy consumption the most and suggest our project partner at Sustainability@BU to take action in order to reach the goal of Carbon Neutral in the future.

Based on historical records and common senses, we made some assumptions that weather contributes the most to both gas and electricity consumption while the effect of other factors remains unknown. So far, after a great amount of effort of data cleaning and method implementing, we have obtained some conclusions which property type takes the most of the effects among all of those factors we listed. On the contrary, the weather actually does not contribute that much as we expected.

Data Description

We came along a distant way of getting feasible datasets and use them. Initially, we planned to use the building energy consumption data directly monitored by the University. However, we were told that the ideal dataset was not accessible to us due to privacy reasons, which contains nearly everything we might need in monthly basis.

In order to get building data of BU, we were suggested to use Building Energy Reporting and Disclosure Ordinance (BERDO) provided by Analyze Boston. Being used as our main dataset, BERDO contains reported buildings' energy usage data and other useful information of the year 2015, 2016 and 2017 in Boston area. To be more specific, in each year's BERDO dataset, it has around 1500 data entries(buildings); each of the entry contains information such as site Energy Use Intensity(EUI), the percentage of electricity/natural gas/steam used, building age, property usage etc. Besides this, there is another minor dataset we made use of, which is monthly average temperatures' data of years 2015, 2016 and 2017 extracted from the website U.S. Climate Data.

The first thing we did is to find BU properties and their corresponding information in the BERDO dataset. In order to achieve this goal, we combined both BERDO Data Map and Boston University Map to select buildings manually. Then we filtered buildings that don't belong to BU. We obtained around 40 buildings' entries in the datasets in three years. Then we are suggested to convert yearly data to a monthly basis by using the distribution of monthly consumption of electricity and gas provided by Sustainability@BU. To achieve this, we did the following:

- We used Site Energy Use Intensity (EUI, unit: kBTU/sq) as total energy consumption containing electricity, gas, and steam(we deleted steam because, according to BERDO, there is only a very small amount of buildings in BU that reported steam percentage)

- Calculate the energy use for electricity and gas by using EUI, the energy consumption percentage of electricity and gas. (Electricity Consumption = EUI * electricity percentage, Gas Consumption = EUI * gas percentage)
- Calculate the monthly energy consumption of electricity and gas by using the distribution provided by Sustainability@BU

After doing the calculation, each building entry in the dataset has 12 more data added to it.

For the next step, we proceeded to the most important part, which is to extract the factors from the BERDO that we would like to inspect and concatenate them with the temperature data we got from U.S. Climate Data. The factors we selected are:

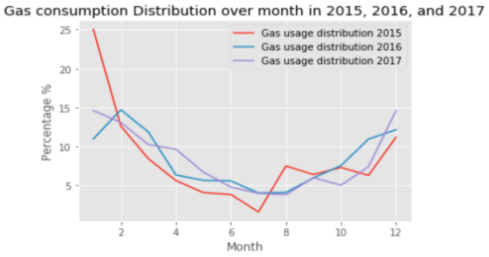
- Monthly Average Temperature in Boston
- Water Intensity (the total water use divided by gross floor area)
- Greenhouse Gas Emission Intensity (greenhouse gas emission divided by gross floor area)
- Building Age (year of the building built)
- Property Type (the purpose of the building)

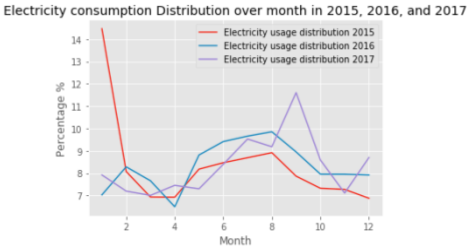
Upon finishing of processing data, we discarded a few buildings that have water intensity and greenhouse gas emission intensity values as None or inf(infinity). Since there are only a few of them, deleting will not make any significant differences in our final results.

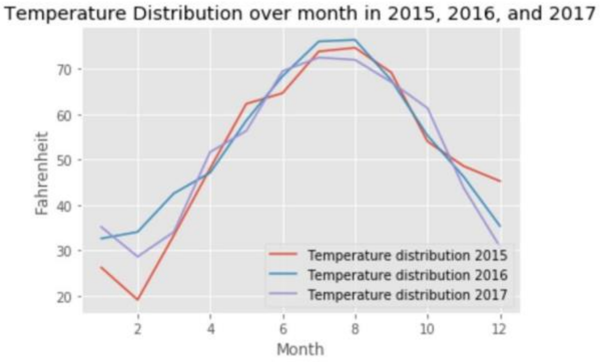
In the end, we stacked up data we processed earlier from the year of 2015 to the year of 2017 and made each data entry under monthly basis so that each building can have multiple entries in the dataset. This can be considered as a data augmentation technique since more data using in our linear model make the model return a more accurate prediction. Finally, we created two files to store the cleaned and ready-to-use data, one for electricity and one for gas, under the names **df_201x_BU_analysis_E_Final.csv** and **df_201x_BU_analysis_G_Final.csv**.

Data Analysis

After exploring our data, we've got several rough visualizations as followings:

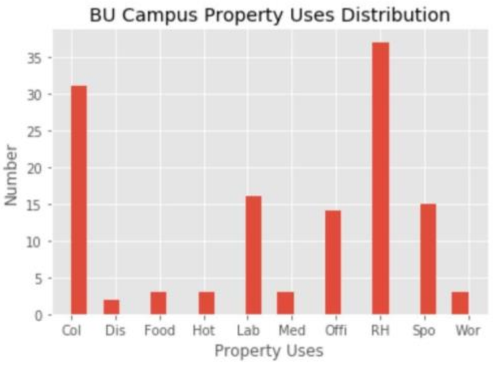


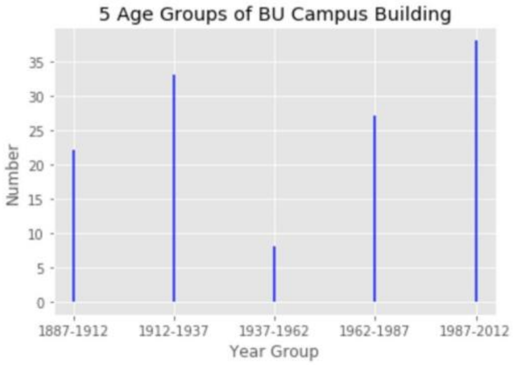




Looking at our visualizations, the first image (upper left) is the distribution of gas usage over months for all of the three years, and the second one (upper right) is the electricity usage distribution over months corresponding to the three years also. From the graph, we can make assumptions that the temperature factor is negatively proportional to the consumption of gas, while, on the other hand, the temperature is positively proportional to the consumption of electricity.

We also visualized the building usage and building age of our dataset, the results are seen as followings:





We defined the types of usage of the BU building in 10 different categories which are:

(Col = College, Dis = Distribution Center, Food = Food Service, Hot = Hotel, Lab = Laboratory, Med = Medical Office, Offi = Office, RH = Resident Halls, Spo = Sport, Wor = Worship Building)

The Algorithm

The approach we used to predict the importance of factors is the standardized linear regression. We want to predict a linear function that can precisely describe the relationship between electricity/gas consumption and the values of potentially important factors listed in the Data Description section. After obtaining the function, we can observe the absolute value of the weight for each factor to determine the importance of it with respect to the electricity/gas consumption.

In order to construct the linear model, we use the following data as factors that potentially contribute to the BU electricity/gas consumption:

- x_{TEMP}: Monthly Average Temperature in Boston
- x_{WI}: Monthly Water Intensity per square feet of the Building
- x_{GHGI}: Monthly Greenhouse Gas Emission Prediction of the building
- x_{AGE}: The Construction Year of the Building

(The following variable are binary variables which indicate whether the building is in a certain type of use)

- x_{COL}: A variable indicating whether the building is for college teaching purpose (classrooms, libraries etc.)
- x_{SPORT}: A variable indicating whether the building is for sports/fitness purpose (stadium, gyms, etc.)
- x_{RES}: A variable indicating whether the building is a university residential hall.

- x_{LAB}: A variable indicating whether the building is a laboratory.
- x_{OFF}: A variable indicating whether the building is a office building.
- x_{HOT}:A variable indicating whether the building is a hotel.
- x_{DIS}: A variable indicating whether the building is a distribution center.
- x_{FOOD}: A variable indicating whether the building is a food service building.
- x_{WOR}: A variable indicating whether the building is a worship/multi-purpose building.
- x_{MED}: A variable indicating whether the building is a medical purpose building.

The reason why we choose ten variables as an indicator vector to express the using purpose of a building is that in this way, after running the regression, we can directly observe which type of usage of the building will impact the electricity/gas consumption and how significant would a specific type impact the energy usage by simple observe the value of its weight, while only using one variable to indicate all the types fails to do that.

After finishing setting up the variables, we define the following two linear functions of electricity and gas consumption:

$$Y'_E = w_0 + w_1 x_{TEMP} + w_2 x_{WI} + w_3 x_{GHGI} + w_4 x_{AGE} + [w_5 x_{COL} + w_6 x_{SPORT} + w_7 x_{RES} + w_8 x_{LAB} + w_9 x_{OFF} + w_{10} x_{HOT} + w_{11} x_{DIS} + w_{12} x_{FOOD} + w_{13} x_{WOR} + w_{14} x_{MED}]$$

$$Y'_G = \theta_0 + \theta_1 x_{TEMP} + \theta_2 x_{WI} + \theta_3 x_{GHGI} + \theta_4 x_{AGE} + [\theta_5 x_{COL} + \theta_6 x_{SPORT} + \theta_7 x_{RES} + \theta_8 x_{LAB} + \theta_9 x_{OFF} + \theta_{10} x_{HOT} + \theta_{11} x_{DIS} + \theta_{12} x_{FOOD} + \theta_{13} x_{WOR} + \theta_{14} x_{MED}]$$

where $Y'_E$is the predicted Site Energy Use Intensity (EUI) of electricity (Unit: kBTU/sf), and $Y'_G$is the predicted Site EUI of gas (Unit: kBTU/sf). Notice that the units of $Y'_E$and Y'_D are kBTU/sqft so that we don’t need to worry about the size of each building which can affect the total amount of energy use for each building.

We want to predict the weights $W = [w_0, \ldots, w_{14}]$ and $\Theta = [\theta_0, \ldots, \theta_{14}]$ of both linear models by using normalized Ordinary least squares Linear Regression. We trained our linear model using 90% of our dataset. Then we used the remaining 10% of our dataset to predict the electricity/gas consumption values and calculate the mean squared error to see whether linear model successfully describe our data.

Data training

To get the predicted$W = [w_0, \ldots, w_{14}]$ and $\Theta = [\theta_0, \ldots, \theta_{14}]$, we calculate the residuals between the function output, Y'_E and Y'_G, and the actual electricity/gas consumption values (EUI), Y_E and Y_G and then find that $W$and Θthat minimize the residuals which are:

$arg\ min_{w}$ || $Y_E - Y'_E$ || for electricity linear model

$arg\ min_{\theta}$ || $Y_G - Y'_G$ || for gas linear model

(||.||is the L2 norm)

In the code, we used the Linear_Model framework from Sklearn to handle the linear regression and calculate the optimized $W = [w_0, \ldots, w_{14}]$ and $\Theta = [\theta_0, \ldots, \theta_{14}]$ for the two linear functions.

Data Testing

By using the approach above, we can obtain the optimal $W = [w_0, \ldots, w_{14}]$ and $\Theta = [\theta_0, \ldots, \theta_{14}]$. Then we use the remaining 10% of our dataset to predict the Y'_E and $Y'_G$and calculate the mean squared errors of two models to see whether the two functions we calculated are suitable to predict the correct electricity/gas consumption EUI.

If the mean squared errors are small, then it means that our model can correctly predict the electricity/gas EUI based on the factors. Also, since our linear regression models are standardized, which means data are normalized before running the regression, we can use the absolute values of the weights to find out which factors play significant roles in energy emission for both electricity and gas.

Experimental Results

After running the linear regression, we obtain the $W = [w_0, \ldots, w_{14}]$ and $\Theta = [\theta_0, \ldots, \theta_{14}]$ and their values are following:

Linear Model for Electricity

Name	Weights	Name	Weights	Name	Weights
TEMP	0.012421	SPORT	298.629369	DIS	298.634887
WI	0.289734	RES	296.960672	FOOD	299.220761
GHGI	3.414407	LAB	302.450182	WOR	297.485089
AGE	0.018668	OFF	299.470882	MED	300.802556
COL	299.319523	HOT	298.134632	CONST	-335.112188

Linear Model for Gas

Name	Weights	Name	Weights	Name	Weights
TEMP	-0.179304	SPORT	-507.408759	DIS	-507.154212
WI	-0.493011	RES	-504.425134	FOOD	-508.270723
GHGI	12.830529	LAB	-513.599862	WOR	-505.363723
AGE	-0.029515	OFF	-508.260151	MED	-510.577937
COL	-508.329972	HOT	-506.639372	CONST	573.183433

And we ran the Mean Squared Error (MSE) test and get:

MSE for Electricity Linear Model	MSE for Gas Linear Model
2.01495783378	17.3252627592

According to the results we obtained above, in the Linear Model for Electricity, the coefficients of TEMP, WI, GHGI , and AGE have very small positive values, and it means that these 4 factors are positively proportional to the electricity energy consumption. The value of type indicator vector shows that LAB factor has the largest value among all of the property types, indicating that if a building functions as a laboratory, its electricity consumption will be higher than other buildings functions as other types. On the contrary, we found that the weights of factor RES has the lowest value among all of the types of usage, indicating that, surprisingly, if a building is a residence hall, then this building will generate less electricity consumption Ceteris Paribus.

In the Linear Model for Gas, we found the results is just opposite to the result of electricity linear model. First, the coefficients of TEMP, WI, and AGE have very small negative values, and it means that these 3 factors are negatively proportional to the gas energy consumption. Also, looking at the type indicator vector, all weights are negative and LAB type building still gives the most impact on the gas consumption among all the types of usage because the absolute value of its weight is relatively larger. Similarly, RES has the lowest impact on the gas consumption, since its absolute value of the weight is relatively smaller.

Conclusions

Looking across our results, although they are not that satisfying as we would expect, we generated conclusions as the following:

- Temperature: Although at first it seems like a big factor that makes a giant impact on energy consumptions, in our model, it is actually not that important as we expected it would be. Also, since temperature is not something that human can control, we would not suggest our project partner worry anything about it.
- Water Intensity: The impact of water intensity exists with the explanation of the possibility of using either electricity or gas to heat water in winter. Based on such results, we would suggest our project partner propose some energy saving campaigns on using hot water.
- Greenhouse Gas Emission Intensity: Compared to other factors, the effect of GHGI is quite large on the perspective of both electricity and gas, in other words, the major amount of electricity and gas consumption at least emits some form of greenhouse gases. Consider the large impact of this factor, we would recommend our project partner to propose the using of clean energy all across the university.
- Age: Age is not quite an important factor in both electricity and gas consumption data. It could be explained by the constant renovation of buildings across the entire BU campus. Thus, we would not suggest anything regarding this factor.
- Property Type: Taking a look at all of those various types of buildings we found, types really is an important factor in the consumption of both electricity and gas consumption. In details, properties that are functioned as laboratories affect either electricity or gas consumption the most. On the contrary, it is surprising to see the residence buildings, either large residence halls or small dormitories, has impacts to energy consumptions least among all types.

Future Steps

Based on our abnormal results, we believe that there still are many deliverables in the future awaiting us to fix and make our results more interpretable and valuable to our project partners. The most important thing is that we have to find out why our linear regression results are so weird that we cannot give any reasonable explanation to it. Based on the improved explanations, we should look out to see if it is possible to implement any other methods that can make our results more explainable such as using p-values or confidence intervals or even improve our linear regression implementation to reach the goal. If any of the deliverables above cannot satisfy our goal of reasonalizing our results, the least thing we should do is to find other factors that might affect either electricity or gas energy consumptions.

Code Link

References

BERDO Dataset: https://data.boston.gov/dataset/building-energy-reporting-and-disclosure-ordinance

BERDO Data Map:

http://boston.maps.arcgis.com/apps/webappviewer/index.html?id=049576c7287f4ee09bcb0a062e43b55c

Boston Weather:
https://www.usclimatedata.com/climate/boston/massachusetts/united-states/usma0046

Boston University Map:
http://www.bu.edu/maps/

SkLearn Linear Regression:
http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

StatsModels Linear Regression:
http://www.statsmodels.org/dev/regression.html

Wikipedia Linear Regression:
https://en.wikipedia.org/wiki/Linear_regression

OLS Regression:
https://en.wikipedia.org/wiki/Ordinary_least_squares

Mean Square Error:
https://en.wikipedia.org/wiki/Mean_squared_error