Prediction of Building Heating Consumption with Easily Accessible Building Information

Wang Xiyu
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA, 15213
+1(412)-452-7673
xiyuw1@andrew.cmu.edu

ABSTRACT
Heating has long been one of the major US home energy use, with 53.1% of the total consumption in 1993, and 41.5% in 2009. Its amount may have significant impact on national carbon emission, greenhouse effect and some other sustainable issues. Thus it is of paramount importance that scientists should look into the factors that may affect heating patterns, and take affirmative actions to reduce, or to optimize. Apart from the living and heating habit of residents, other factors that can affect the amount of energy in heating might be the building itself. Therefore, this paper looked into some easily accessible factors, namely, number of heated rooms, heated areas, built years, and the region where a building stands. This paper then applied data-driven approaches to find the correlation and make predictions when necessary. The model and the conclusion of this paper can be applied in future design process, to determine the most energy-efficient building type in a particular area.

CCS Concepts
• Computing methodologies → Linear regression → Regression tree and Prediction.

Keywords
Heating consumption; Prediction; Regression; Heated area; Number of heated rooms; Built year; Region; Data driven approaches.

1. INTRODUCTION
Heating consumes a large portion of the energy use in an average household (with a mean of 36863 BTU) across the whole nation. Several parameters that can affect heating consumption are related to personal habit and preference, such as heating degrees, heating during peak hours. Other factors concerning the building type itself can also positively influence the energy efficiency of a building, and its heating consumption in the very end. Intuitively, such factors may include built years, number of heated rooms, heated areas, and the region where a building stands, just to name a few. By understanding the impact these features can exert on the consumption pattern [1], engineers could tailor and design future buildings to make them most energy-efficient without disturbing the living styles of their residents.

1.1 Proposed Approach
Despite of the discrepancies in daily, or monthly heating consumption, the yearly consumption is much more robust and valuable in the perspective of natural preserving, therefore, this paper only addresses yearly consumption data.

This paper first conducts exploratory data analysis and visualization to achieve a preliminary understanding of the data. A very basic assumption about the data could be proved in this way. Then this paper takes commonly applied data analysis methods to study the topic brought about in the introduction part.

To understand the influence of the aforementioned building information, this paper adopts a decision tree, uses those factors as features, and takes the heating consumption as the result. By training the decision tree, a model can be achieved and can then be applied to test data to make predictions. The essence of this process lies in tweaking the features, and mapping high dimensional data to 3D plots. Detailed explanation about the results of each regression is provided.

1.2 Dataset
The source of the dataset is Residential Energy Consumption Survey (RECS) [2] that provides energy related data with multiple characteristics in a 5-year interval. The data used in this paper is collected in year 2009. Data collected in 2005 or before could also be adopted, but there exist errors concerning the heating consumption in those datasets. This dataset is not rather large compared with other sorts, which renders it appropriate for a course project such like this, but robust and stability is the price paid for such convenience.

In the codebook of this dataset, the total usage for space heating in BTU is referred to as “TOTALBTUSPH”, built year as “YEARMADE”, total heated area in square footage as “TOTHSQFT”, number of heated rooms as “HEATROOM”, and the region of the building location as “REGIONC”, in which the number 1-4 correspond to Northeast, Midwest, South, and West respectively.

Other characteristics not related to this topic can be dropped to make the analysis more efficient.

2. RESULT
2.1 Exploratory Data Analysis
Common sense has informed us that heating consumption is more or less related to the regional location of a building. To confirm this sense, the dataset adopted in this paper is classified by the region code, and then four sets of heating consumption over the built year of a building is plotted on the same figure, shown in Figure 1.

Despite of the relatively small amount of samples concerning buildings built in 1920-1950 period, the energy consumption of the buildings seem to follow a nationwide trend. And it can also be inferred from the figure that there are several spikes when the building was built in 1920, 1930, 1940, 1950 and so forth. The scatter plot is quite clear about the distribution of heating consumption for a given built year.

To look into the amplitude of the consumption, and within each region, stem plots can be applied (see Figure 2).

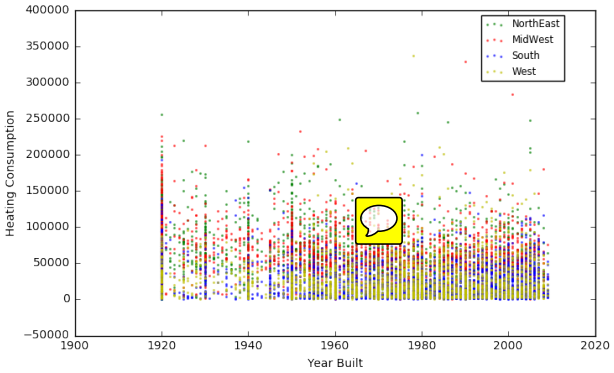


Figure 1. Heating consumption of buildings constructed in 1920 – 2010.

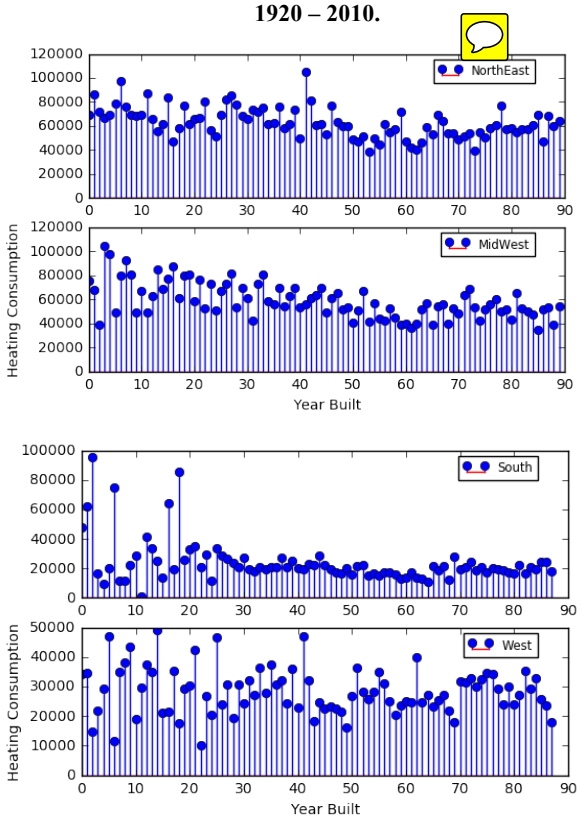


Figure 2. Heating consumption of buildings built in 1920-2010 in four regions.

Despite of several outliers, it is clear that the heating consumption in South US is much smaller in amplitude compared with other regions, which is in accordance with our common sense that the south of US is far warmer than other parts.

2.2 Relationship among Yearly Heating Consumption, Region, Heated Area, and Number of Heated Rooms

2.2.1 Optimized number of heated rooms

It is expected that when the number of heated rooms increases, heating consumption is also increasing, but when that number reaches a certain threshold, the increase in the heating consumption will become very slow, and that the consumption might remain at the same level afterwards. This phenomenon may due to convection, radiation and many other factors. Then for buildings in a particular region, optimized number of heated rooms may exist. This preliminary thought could be indicated from the following boxplot (see Figure 3).

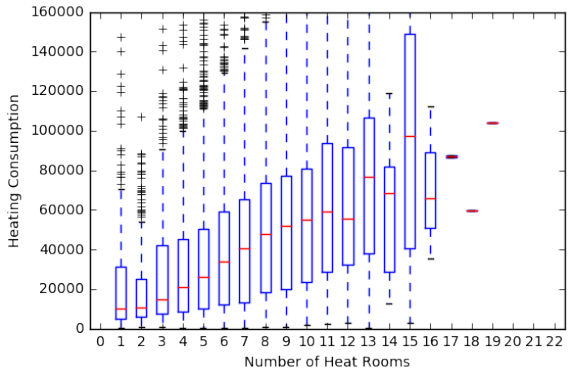


Figure 3. Relationship between heating consumption and the number of heated rooms.

As can be inferred from the above figure, when the number of heated rooms reaches 8, the pace of the increase in heating consumption becomes very slow. The yearly consumptions of 9 rooms, 10 rooms, and possibly 12 rooms are almost at the same level. In that case, building a 10-room household instead of an 11-room one for instance, is not as significantly more economic as we thought might be.

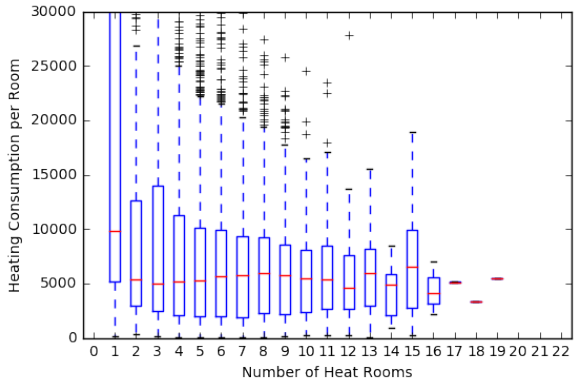


Figure 4. Tendency of heating consumption per room as room number increases.

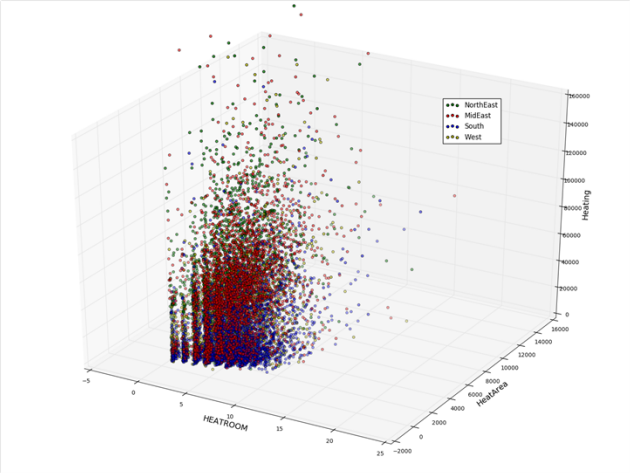
The heating consumption per room is shown above. It can be inferred that building with only one room is least favourable, since it suffers the highest consumption rate. The rate drops when the room number is 2-3. And it increases from 3-8, and then drops considerably from 8 to 12. It remains almost the same at 14, and 16. So the optimized heated room number might be somewhere around 3 and 12, specially 3 for small households, and 12 for large apartments.

However, due to the small amount of data in this dataset, the result is also not very stable. Outliers exist, which intervenes my judgment about the result. Also, the small amount of data prevents me from conducting analysis according to the region code. Otherwise, in some cases, I will have less than 10 data in a particular region about a particular heated room number.

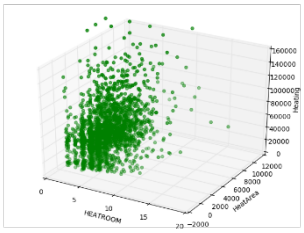
2.2.2 Decision Tree Regression

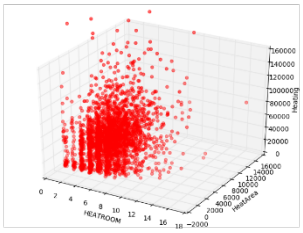
A decision tree was also fitted to the dataset. When a regression tree is fitted to the whole dataset, factors such as the number of

heated rooms, heated area, and region should be regarded as three features. This is because the fitted tree should be able to predict the heating consumption if those factors are given. To fit the regression tree, a library in python named as sklearn is adopted. Since heated area is discrete, the depth of the regression should be restricted to prevent overfitting. Result of this regression is plotted by Axes3D in python. And the 4th dimension, namely, the region code, is distinguished by color (see Figure 5).

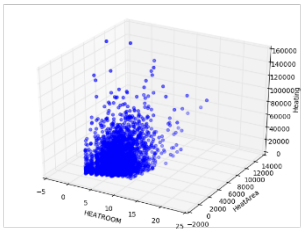


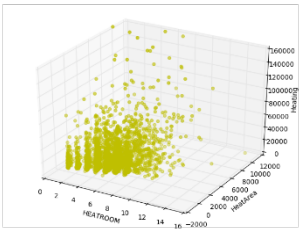
(a) Regression to the whole dataset





(b) Regression in Northeast **(c) Regression in Midwest**





(c) Regression in South **(e)Regression in West**

Figure 5. Result of the decision tree regression.

This does not achieve what I expected in the first place. There should be stronger correlation between the number of heated rooms and heating consumption than the regression shows. Therefore, the decision tree regression method is rather unsuitable for predicting this relationship.

2.3 Relationship among Yearly Heating Consumption, Region, Heated Area, and Built Year

Built year is an indirect indication of a building's condition, thus it may have some correlation with the yearly heating consumption of a building.

2.3.1 Decision Tree Regression

The basic procedure is similar to the above section. The main difference is that instead of using the unsatisfactory feature "number of heated rooms", built year is adopted here. So in this regression, when a regression tree is fitted to the whole dataset, factors such as built year, heated area, and region should be regarded as three features. In this way, the fitted tree would be able to predict the heating consumption if those factors are given.

2.3.2 Cross Validation and Grid Search

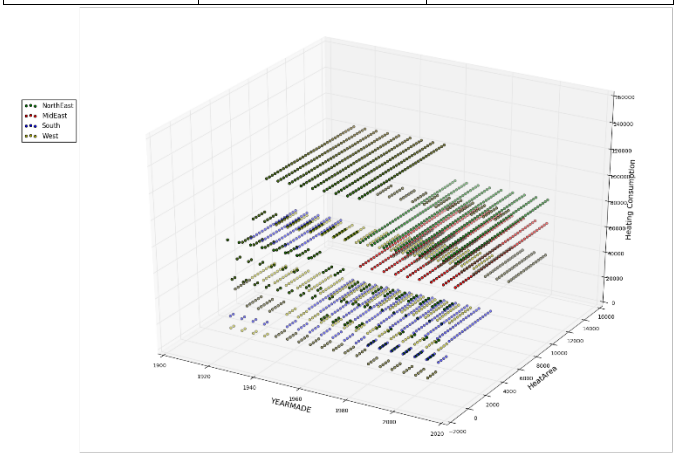
A 10-fold cross-validation is applied, which randomly separates the dataset into 10 folds. Since the regression is subject to overfitting, its depth should be restricted. To find the best depth, grid search was adopted, and the depth with highest cross validation score is considered as the optimum depth. And the corresponding model is stored for the following prediction.

2.3.3 Prediction

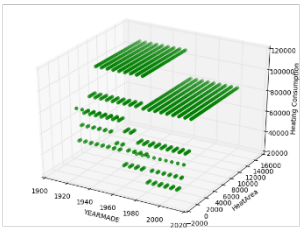
A dataframe of all the possible combinations is created as a test dataset. The range of each feature is listed in Table 1. Heating consumption with respect to each combination is predicted, and is ploted in Figure 6.

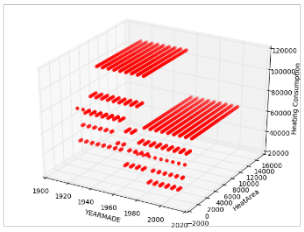
Table 1. Range of three features: Region, Built Year, and Heating Area.

Region	Built Year	Heating Area (square feet)
1-4	1920-2010, with a 5-year interval	100-14000, with a 300 interval

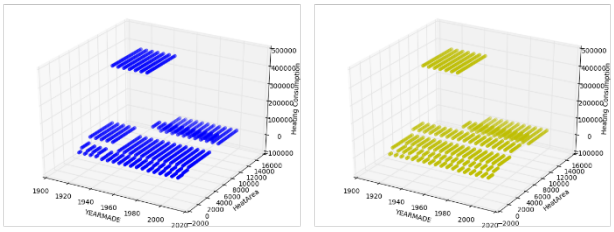


(a) Prediction to the whole dataset





(b) Prediction in Northeast **(c) Prediction in Midwest**



(c) Prediction in South **(e) Prediction in West**

Figure 6. Result of the decision tree regression.

This prediction yields some interesting facts:

(1). Generally, heating consumption increases as heated area increases. The reason why this relationship is discrete is that I created discrete areas with a gap of 300, rather than making it continuous. And the highest platform may due to less training data within that range.

(2). There is a huge gap in heating consumption between houses built before 1960, and after 1960. It is obvious that houses built after 1960 are much more energy-efficient. And it seems to be a nationwide trend. This sharp drop in heating consumption might result from a revolution of building standards, or the wide acceptance of a new material. Either fact is true, engineers should seek to retrofit those old buildings with that measure, and also considerate it when designing a new building.

3. CONCLUSION

This paper tentatively explores the relationship between heating consumption, and easily accessible building information such as number of heated rooms, built year, heated area and regional location. The conclusions of this paper are:

(1) The yearly consumptions of 9 rooms, 10 rooms, and possibly 12 rooms are almost at the same level.

(2) Building a 10-room household instead of an 11-room household is not as significantly more economic as we thought might be.

(3) The optimized heated room number might be somewhere around 3 and 12, specially 3 for small households, and 12 for large apartments.

(4) The decision tree regression method is rather unsuitable for predicting the relationship among heating consumption, number of heated rooms, heated area and region.

(5) There exist certain correlation among heating consumption, built year, heated area and region.

(6) Generally, heating consumption increases as heated area increases.

(7) There is a huge gap in heating consumption between houses built before 1960, and after 1960. And it seems to be a nationwide trend.

(8) When designing a new building, engineers should consider the optimized number of rooms, heated area, and the cause for the drop of heating consumption concerning buildings built after 1960. An energy-efficient building can be designed in this way.

4. FUTURE WORK

As stated above, this paper only looks into the data in 2009, which may not be a typical representative of heating consumption data across the whole time series. Further study should involve importing more datasets to validate the conclusions in this paper, as well as making more solid judgment about the optimized room number. Also, a new regression model other than decision tree should be adopted to improve the prediction of the relationship among heating consumption, number of heated rooms, heated area and region, so that the fitted model should be able to predict the heating consumption if those factors are given.

5. REFERENCES

[1] Reddy, T. Agami. 2011. Applied data analysis and modeling for energy engineers and scientists. *Springer Science & Business Media.*

[2] U.S. Energy Information Administration - EIA - Independent Statistics and Analysis. (n.d.). Retrieved December 10, 2016, from https://www.eia.gov/consumption/residential/data/2009/