Analyzing Fuel Oil Consumption in U.S. Households Census by Region and Division

Keyi Huang (huang1), Yue Pan (ypan1)

ABSTRACT

Energy consumption in households in U.S. is crucial for policy decision make, planning energy and sustainable development. Since energy consumption in households is very complex and location-specific due to the different local climates, policies, areas and etc., the patterns of the consumption in households may vary from region to region. So we conduct an analysis of the U.S. household fuel oil consumption data to gain insights into the oil consumption of different regions and the nation as a whole.

We build clustering model and regression model to present analysis of the patterns and trend of fuel oil consumption by regions and divisions from 1980 to 2001. Total nine different regions (New England, Middle Atlantic, East North Central, West north Central, South Atlantic, East South Central, West South Central, Mountain and Pacific) have been analyzed and classified into four clusters just corresponding to four divided geographical parts of U.S. (Northeast, Midwest, South and West). That is, the datasets successfully indicate that the consumption patterns vary from region to region. We provide some related potential reasons contributed to this regional difference distribution. Finally, we study some factors affecting the total oil consumption over time and our regression model predicts that there is a long-term downward trend of oil consumption in nation-scale.

CCS Concept
•Information systems → Data analytics; Clustering; Regression

Keywords
Fuel Oil Consumption, Cluster Analysis, Regression Analysis

1. INTRODUCTION

1.1 Motivation

In the U.S., most of the energy comes from nonrenewable energy sources. Coal, petroleum, natural gas, and uranium are examples of nonrenewable energy sources. Nonrenewable energy sources are used to make electricity, to heat our homes, to move our cars, and to manufacture products [1]. And fuel oil, gasoline and other liquids produced refined from petroleum provide people with use for many different purposes in daily life.

So it is of great importance to figure out the household energy consumption characteristics in nation-scale, which has been done partly with the distribution across the nation by regions and the trend of the total consumption in our study. A better understanding of the household energy consumption patterns and trend would help the decision maker improve the aggregation and organization of the energy structure, which finally could optimum utilization of energy resources and reduce the wastage of energy.

1.2 Background

The United States consumes more energy from petroleum than from any other energy source and among the petroleum usages, fuel oil accounts for the largest portion [2]. So we focus on the analysis on the fuel oil consumption in U.S. households and try to understand the types of the consumption pattern.

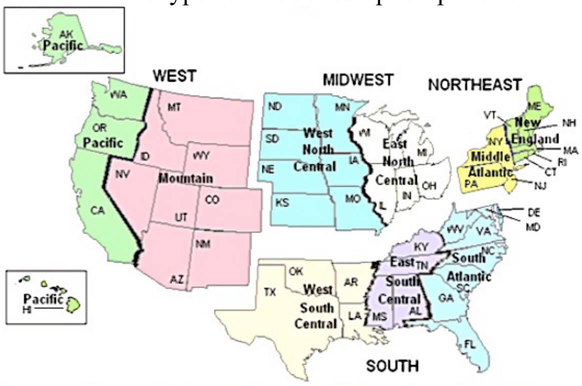


Figure 1: U.S. regions and divisions [3]

Figure1 shows the map of U.S. divided by four main regions: Northeast, Midwest, South and West. They all have some unique characteristics and the differences among them allow us to study the diversity of fuel oil consumption patterns.

1.3 Related Work

There are several tasks we need to do:

1. Figure out the distribution curves of oil consumption in U.S. households across different regions from 1980 to 2001.
2. Try to answer how many types of curves exist with the dataset and why.
3. Find out how the total oil consumption is influenced by time, household characteristics, weather, oil price and some other factors.
4. Try to build the regression model with a better prediction on certain selected predictors (i.e. features).

2. PROPOSED APPROACH

We study the regional household fuel oil consumption patterns and trend by clustering analysis and regression analysis. And also we analyze some factors that might influence the oil consumption over last decade.

2.1 Clustering Analysis

We use the k-means clustering algorithm to analyze the data of oil consumption per building in different divisions of the region. K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem [4]. The procedure follows a simple and easy way to classify a given dataset. It aims to partition the input data set into k partitions (i.e. clusters). In this problem, we use k-means clustering to partition nine divisions (these are observations) into four clusters in which each division belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

One of the advantages of this algorithm is easy to understand and it can give the best result when data set are distinct or well separated from each other. Also, it's relatively efficient to divide a huge amount of data into only several clusters. In this case, we divide nine data into four clusters. Although it's not as effective as normal, it can still reveal the result of how the division of regions should be fitting into the model.

2.2 Regression Analysis

In doing the regression analysis, we build the prediction model with machine learning algorithm, decision tree. Decision tree builds regression models in the form of a tree structure. It would break down our dataset into smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches, each representing values for the attribute tested. Leaf node represents a decision on the numerical target [5]. In our case, we need to test and fit a regression tree to the dataset we got and predict the total fuel oil consumption.

When doing the feature selection and model optimization, we keep calculating the score for every regression model we built. Score is also defined as the coefficient of determination R square of the prediction. A score that is quite close to 1.0 means that the model has an accurate prediction on the response variable, in our case, the total oil consumption. Also we keep monitoring the feature importance after we build each model. It is also known as Gini importance that is more likely to be criteria. The higher feature importance, the more important feature [6]. The threshold value we would be using is for feature selection, too. Features whose importance is greater or equal are kept while the others are discarded. In our work, we choose "median" (i.e. the median value of all of the feature importance) to be the threshold.

3. DATASET

The dataset used in this study was collected from Residential Energy Consumption Survey (RECS) offered by U.S. Energy Information Administration (EIA) and we gathered the data from year of 1980 to 2001.

Num. of residential households	15.4 million
Num. of residential buildings	11.6 million
Total floor space of the residential building	29.7 billion sq. ft.
Oil price	116.72 dollars per barrel
Fuel oil consumption (divided by nine regions)	169.0 trillion Btu
Duration	21 years

Table 1: Key attributes of the dataset

This dataset could answer various questions on U.S. household fuel oil consumption patterns and trend. We build the clustering model with nine different regions data, and take the total fuel oil consumption as the label (i.e. response) and other characteristics as the features (i.e. predictors).

4. RESULTS

4.1 Data analysis

4.1.1 Boxplot of oil consumption per-building

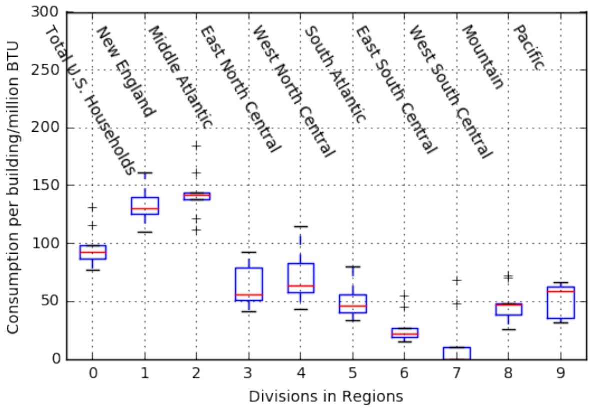


Figure 2: Boxplot of oil consumption per-building in different regions of divisions

Boxplots are very good at presenting information about the central tendency, symmetry and skew, as well as outliers, although they can be misleading about aspects such as multimodality [7]. One of the best uses of boxplots is in the form of side-by-side boxplots, just as Figure 2 shown. It contains the full range of oil consumption (from min to max), the likely range of oil consumption and the median in every division from 1980 to 2001. Figure 2 also display surprisingly high maximums or surprisingly low minimums that are the outliers in the datasets we use.

From the boxplot in Figure 2, we can get that the variation range, median and outliers of oil consumption in the nine census divisions from 1980 to 2001. The variation range of Middle Atlantic and East South Central is much smaller than other seven divisions, which means annual oil consumption in the two divisions almost keeps the similar amount. However, change range of East North Central, West North Central and Pacific is more than five times of the change in Middle Atlantic and East South Central.

And some data of several regions have some outlier data which might be mistakes or otherwise unusual. There are more outliers in Middle Atlantic than others. In addition, we can find outliers in East South Central and West South Central easily from the boxplot. The more the amount of outliers is, the more unstable and irregular the data is. Oil consumptions in New England and Middle Atlantic within past 20 years are always higher than oil consumption in other region, so that we can assume that people in New England and Middle Atlantic depend on oil in a greater extent. On the contrary, West South Central and East South Central consume less oil than other regions.

4.1.2 Heat map of oil consumption per-building

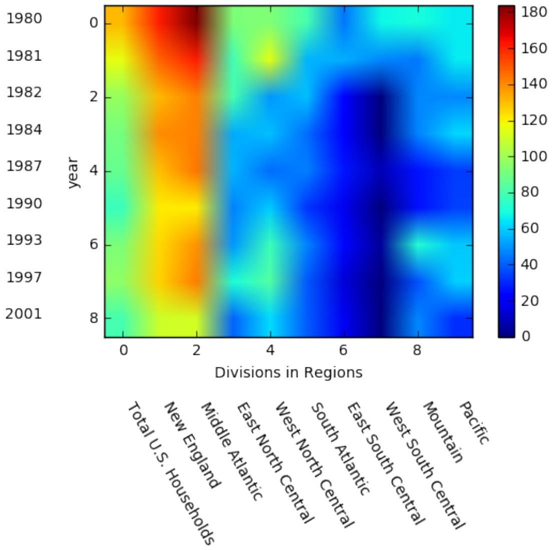


Figure 3: Heat map of oil consumption per-building in different regions of divisions

A heat map is a graphical representation of data where the individual values contained in a matrix are represented as colors, just as Figure 3 shown. The color bar on the left hand shows the variation of oil consumption with color changing. As a result, Figure3 displays the oil consumption change trend directly by the color change through year 1980-2001 in the vertical direction. It is an intuitive approach to express consumption variation.

Through variation of color in Figure 3, we can see directly that the oil consumption in almost all divisions have been reduced gradually with the time passing. However, some data do not follow this rule restrict, which means there are some outliers. Temporary increased occurred in some census divisions, but no evidence points to a reversal of the long-term declining trend.

What's more, color in East South Central and West South Central is nearly deep blue which means that these two areas consume less oil. Bright color appears in New England and Middle Atlantic so that oil consumption in these areas is larger than others. It also verifies the results we get from Figure 2.

4.2 Clustering

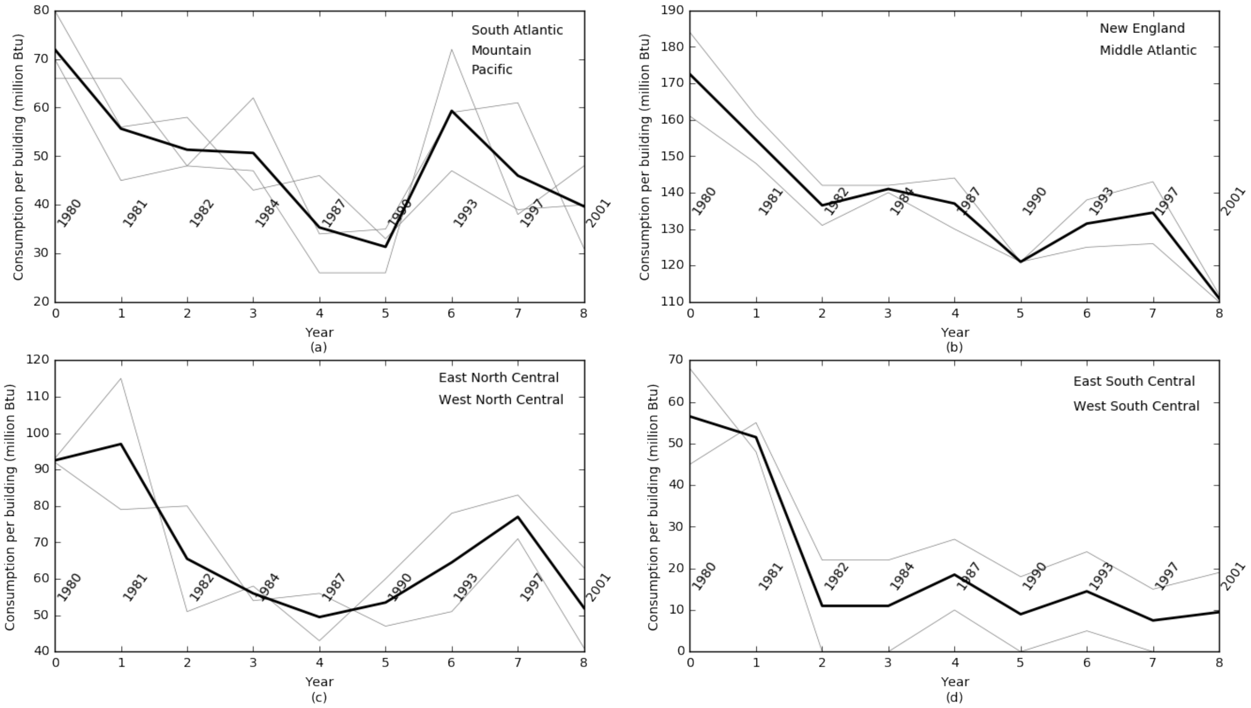


Figure 4: Clusters (a) to (d) show four different of fuel oil consumption patterns, analyzed by nine different regions.

Cluster	Region	True Value	Observed Value	Accuracy
a	West	Mountain, Pacific	South Atlantic, Mountain, Pacific	50%
b	Northeast	New England, Middle Atlantic	New England, Middle Atlantic	100%
c	South	South Atlantic, East North Central West North Central	East North Central West North Central	67%
d	Midwest	East South Central West South Central	East South Central West South Central	100%

Table 2: Clustering results

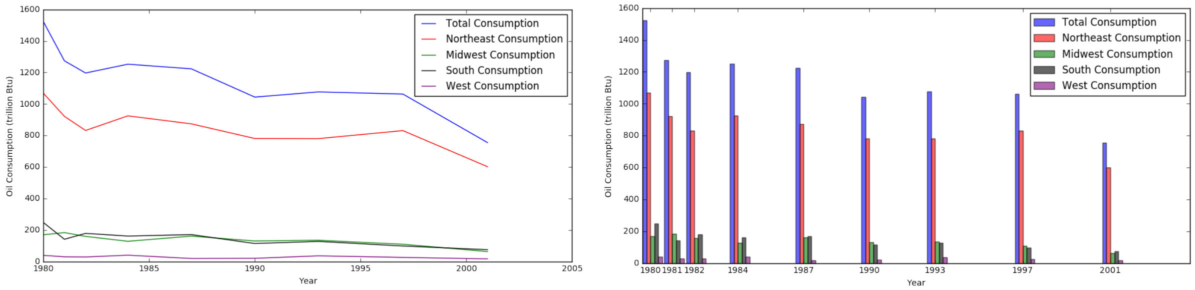
Oil consumption per building can be divided into four clusters by census divisions according to the k-means method, as Figure4 shown.

Census divisions in the same region almost have the similar oil consumption per-building variation tendency. Most of the way of how the data is divided into clusters is the same as the way regions West, Northeast, South and Midwest are divided. It is worth noticing that the division South Atlantic is an exception. In fact, South Atlantic belongs to South region, but it has similar oil consumption varying trend to West region. What's more, South Atlantic uses less oil than other two divisions in South region.

From 1980 to 1982, nine census divisions show sharp decreases in oil consumption per building. The largest percentage decline in per-building consumption occurred in the West South Central and West South Central census division. However, these two regions exhibited relatively stable trend after 1982. From 1990 to 1993, oil consumption increase in all nine divisions. And South Atlantic, Mountain and Pacific division see the rapider rise than other six divisions. That is to say, we can see that the data in the same clusters have the similar varying trend during these years, which means the different clusters have different varying trends. But there is still some similarity between different clusters. Plus, the average value of data in one cluster is obviously different form that in another cluster.

4.3 Regression

Given the four clusters of the fuel oil consumption patterns, we plot the consumption curves separately by four main regions as well as the total consumption.



(a) Line plot (b) Bar chart

Figure 5: Fuel Oil Consumption Census by Region and Division, 1980-2001

Model	Importance and selection of Features X					Max depth	Score
	Year	Num. of households	Oil Price	Num. of buildings	Building Area		
1	1.0000 Y	0.0000 Y	0.0000 Y	0.0000 Y	0.0000 Y	1	0.6025
2	0.8067 Y	0.0000 N	0.1881 Y	0.0000 N	0.0053 Y	3	0.9953
3	0.8037 Y	0.0037 Y	0.1874 Y	0.0000 N	0.0053 N	4	0.9990
4	0.8067	0.0053	0.1880	\	0.0053	3	0.9953

Table 3: Feature selection in regression model

Figure 5(a) and (b) show the fuel oil consumption curves over 21 years. A long-term trend in declining U.S. household fuel oil consumption is apparent. Then we build the regression model with label Y (i.e. total consumption in nation-scale) and all possible features X (year, number of households, oil price, number of buildings and the area of the floor space). We optimize our regression model by selecting the features and keeping calculating and improving the score. Table 3 shows the process of selecting the features in the regression model. Note that "Y" means that it is suggested to keep that feature and "N" means that we could get rid of feature in the model. When set up the threshold to be the median of all of the importance of features, we could select the features with relatively high significance (i.e. importance). By evaluating the feature significance, together with a relatively high score of the model, we finally choose the regression model with the features of ***Year, Number of households, Oil price and Building area*** (get rid of the feature ***Number of buildings***).

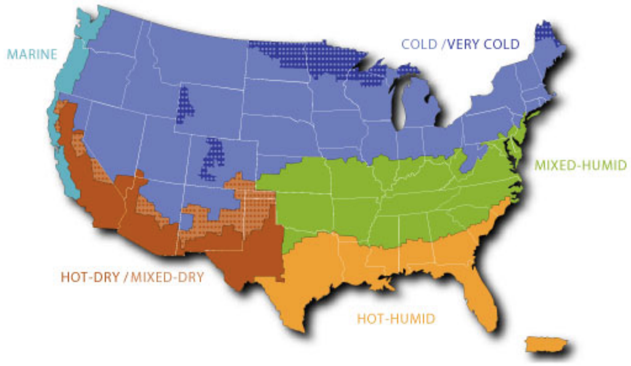


Figure 6: U.S. climate regions map [8]

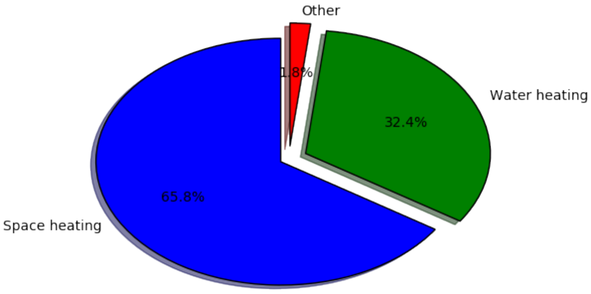


Figure 7: Oil consumption in homes by end uses

5. DISCUSSION

We classify total nine census divisions into four clusters (only one region ***South Atlantic*** falls into a wrong region where it does not locate in), which indicates that the dataset we collect is to some degree significant to show the different patterns of oil consumption in different U.S. region. We can do some reasonable assumptions to explain why South Atlantic does not belong to South region:

For instance, both of the ratios of the length of the regional boundary along the seashore to the area of the whole region are relatively significant. So they may have a natural advantage of trading through the sea. As a result, it makes sense that they may have the similar development model, which means they may use the similar energy to complete the requirement of the whole divisions.

Also, there are many world-known tourist cities in the South Atlantic area, such as Orlando, Miami and Panama City. So they need to find more environment friendly energy to keep the city beautiful and attractive. As a consequence, they may use other clean energy to replace oil and have lower oil consumption than other cities nearby in the South region.

The long-term downward trend of the total consumption curve may due to lots of reasons. We finally select four features (influencing factors) with relatively high significance by building the regression tree. There are more facts that may have affected the total consumption to decline:

Noted that the Northeast always accounts for the largest part of the oil consumption, it may be resulted from the climate (i.e. the temperature). Compared with other three regions, the Northeast region has colder weather during the year. And Figure 7 shows that 65.8% usage of fuel oil in household goes into the space heating, which means the colder region might need to consume more fuel oil to warm up the house, which also makes sense that why the total oil consumption is to some degree related to the building area as the regression model shows. So temperature might play an important role in fuel oil consumption, especially in the Northeast. And the change of the temperature in every region may result in the change of total fuel oil consumption.

Also there is another fact that the energy consumption structure has been always changed and reunited. Actually only a small amount of crude oil is directly consumed in the United States. Nearly all of the crude oil that is produced in or imported into the United States is refined into petroleum products such as gasoline, diesel fuel, heating oil, and jet fuel, which are then consumed. But these are all nonrenewable energy resources. Now some of the renewable energy sources are also used for electricity generation, heat generation, and transportation fuels. Some renewable biofuels, such as ethanol and biodiesel, are used as substitutes for or as additives to refined petroleum products. So this may contribute to the decline of the total fuel oil consumption in nation-scale.

6. FUTURE WORK

There are a few things we could do to further improve our work on analyzing the fuel oil consumption in U.S. households:

As for the dataset, we should collect more data with time series (e.g recorded data in every month of a year), and collect the data in recent years. By this more detailed dataset, we could figure out more about the seasonal characteristics in the oil consumption patterns and predict the near future trend of the total consumption in nation-scale.

Also, since four main regions (Northeast, Midwest, South and West) show four different types of the curves, more features could be taken into consideration when building the regression model (e.g the temperature, population, etc.) to make the model have a better prediction on total fuel oil consumption.

7. REFERENCES

[1] Energy Explained: Crude Oil and Petroleum Products–Use of Oil
[2] The Energy Information Administration's (EIA) Crude Oil Inventories
[3] Energy Information Administration, Short-Term Energy Outlook
[4] Tapas Kanungo, David M. Mount. An Efficient k-means Clustering Algorithm: Analysis and Implementation.2002
[5] Jason Brownlee. Classification And Regression Trees for Machine Learning. 2014
[6] Menze BH. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. 2009
[7] Benjamini Y. Opening the Box of a Boxplot. 1988
[8] Residential Energy Consumption Survey (RECS) U.S. Climate Regions Map. 2009