Linear Regression Model of Individual Electric Household Power Consumption in France

Jue Hou
Civil and Environmental Engineering Department
jueh@andrew.cmu.edu

Yeze Wang
Civil and Environmental Engineering Department
yezew@andrew.cmu.edu

Yue Zhang
Civil and Environmental Engineering Department
yuez4@andrew.cmu.edu

Yunyang Liu
Civil and Environmental Engineering Department
yunyangl@andrew.cmu.edu

ABSTRACT

The household energy consumption is always the vital concern for both the families and government. This project is aimed to find an efficient way to discover the possible correlation between energy consumption and the holiday. Through data mining, clustering and linear regression, the household energy consumption, temperature and holiday data are analyzed and the process and results shows are shown in the article below.

KEYWORDS

Energy Consumption, Temperature, Clustering, Linear Regression

1 INTRODUCTION

On holidays, most American families choose to spend their free moments out instead of staying at home. This may influence the household energy consumption during holidays. The project is aimed to conduct a series of analysis on the relationship between holidays and household energy consumption. Including the comparison of energy consumption between workdays and holidays, the difference of staying at home and going on a trip during holidays, the effect of the length of holidays have on the household energy consumption.

The data used included the household energy consumption data, the temperature data and the holiday data.

The method we will be using is first converting the text data into a csv file to make it easier for the next steps. Then we will plot figures of the raw data to try to find some regulations of the household consumption during the period, including the consumption curve of each year as well as the curve of a single day. Through this we can find out the peaks and troughs of household energy use. Then we will conduction clustering to the energy consumption data to classify the holidays and workdays. Through comparing the consumption of the two classes, find out the relationship between energy consumption and holidays. After that, we can conduct linear regression to find out the length of holidays and the energy consumption to seek the influence the lengths of holidays have on the energy use.

The way we will be using to validate our project is to use other datasets of household energy consumption o prove the conclusion in our project is right. If the relationship we find is also right in other household's dataset, it shows the result of our project is convincing.

2 Data Description

2.1 Household Energy Consumption Data

The energy consumption data used is the individual household electric power consumption dataset released by EDF Energy. It is the data of energy measurements of a single household in France from 12/2006 to 11/2010. Its measure resolution is 1 minutes, it has the active and reactive power, voltage, and 3 sub-meter data. The dataset can be found at the UCI Machine Learning Repository [1]. The link is in the reference below.

2.2 Household Temperature Data

The temperature data is the average temperature of Paris, France The temperature data range is also from 12/2006 to 11/2010, the same with that of consumption data. Its data resolution is 1 hour. The data can be found at the MeteoBlue website[2].

2.2 Holiday Data

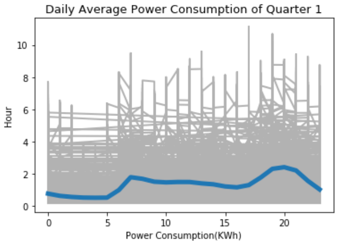
The holiday data is an indicator that shows whether the day is holiday or not. The holiday are found according to the France national holidays, which can be found at Timeanddata.com[3].

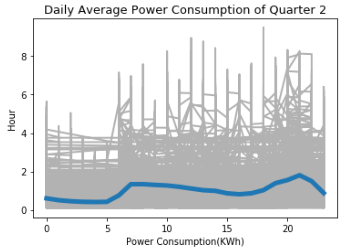
3 APPROACHES AND RESULTS

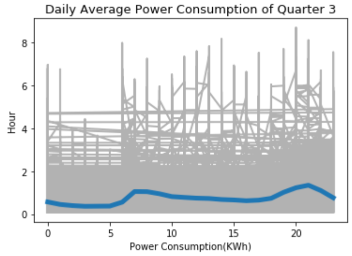
3.1 General Analysis Of Data

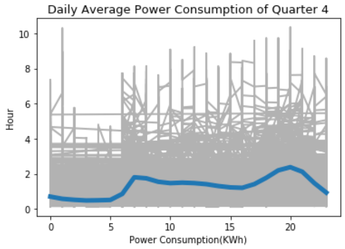
Divide the data into 4 quarters of the year, it can be seen that the mean of average power consumption of this individual household varys from season to season. During summer the average energy consumption is the lowest and during winter is highest. It can be inferred that this household needn't spend energy on cooling during hot weather while it uses much more energy on heating during cold days. The pulse in the data may due to the sudden activation of refrigerator. The box plots indicates that there are certain amount outlier data.However, considering a 4-year-period observation, those can be neglected.

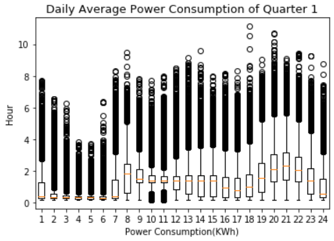
There is a significant rise at 6:00 am and 6:00 pm while slowly decrease at 8:00 am and 11:00 pm. It can be deducted that the sleep hours and work hours has a significant influence on power consumption.

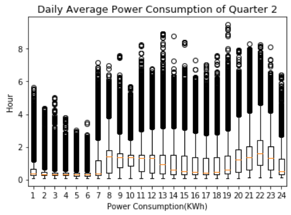


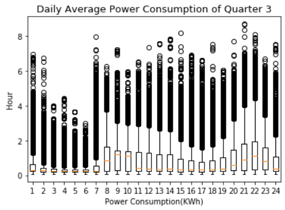


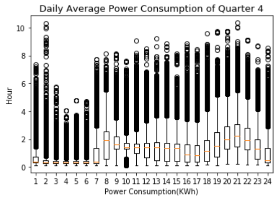










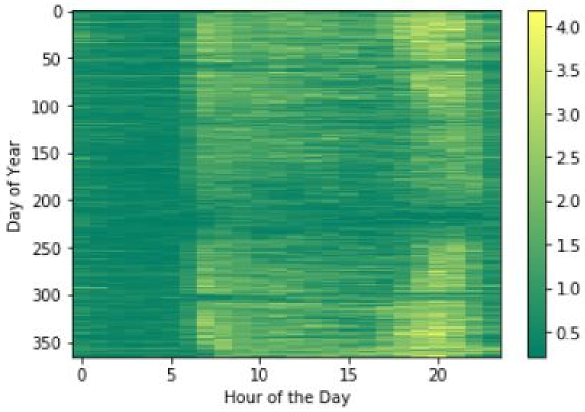


Thus, it can be assumed that the power consumption is sensitive to temperature and whether it is in work hours or sleep hours. To build the model, those values can be considered as variables. Combine the temperature, national holidays and the power consumption data, a linear regression model can be conducted.

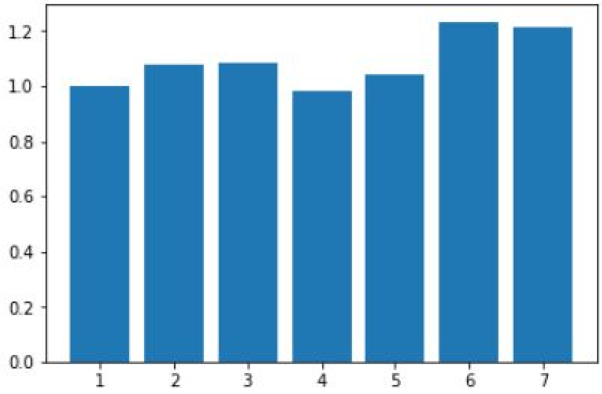
3.2 Visual analysis of annual and weekly electric power consumption pattern

In order to further investigate how household electric power consumption would change according to different parameters, annual and weekly analysis is performed. For the annual power consumption analysis, load curves that consist of each day in a year is created based on the electric power consumption hourly. The the load curves are shown as a density map to indicate the power consumption pattern annually.

As it is shown on the image, lighter colors represent higher electric power consumption. There is a obvious trend that from 0 am to 6 am, the electricity consumption is low, which is reasonable since people are sleeping at that time. The electricity consumption increases in daytime with two separate peaks, which take places in the morning and around 8 pm. The morning peak means that people are getting up and preparing to work. The evening peak represents the time when people are coming back home from work.



The figure below shows weekly pattern this household power consumption. The bar plot shows the electric power consumption on each day of a week. It can be observed that electric power consumption is slightly higher on weekends than on weekdays, which make sense since people would have more time spending at home.

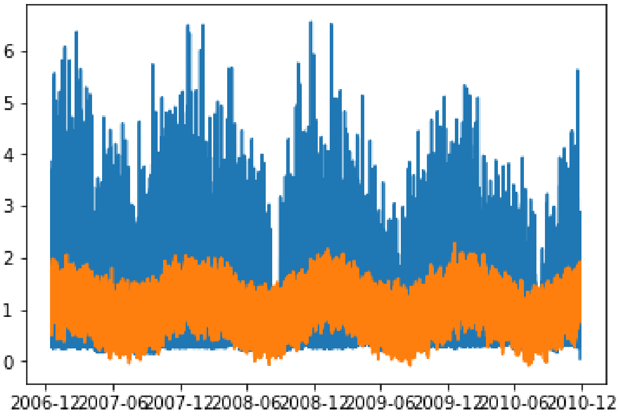


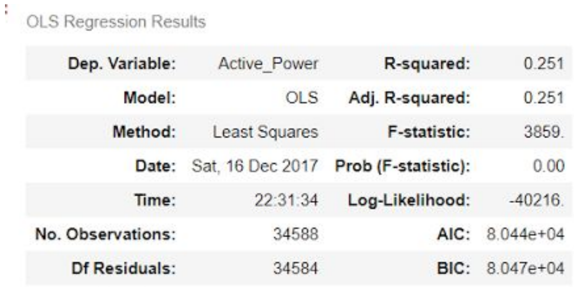
3.3 Linear Regression Model

The equation of regression model is as follow:

$$Consumption = \alpha + \beta Temperature + \gamma(Work\ Hour) + \delta(Sleep\ Hour)$$

The work hour and sleep hour are booleans. Work hour is ranged from 9:00 to 17:00 and sleep hour is ranged from 11:00 to 6:00. When it is a national holiday or weekend, work hour will always be 0. Meanwhile, sleep hour remains the same.





OLS Regression Results

Dep. Variable:	Active_Power	**R-squared:**	0.251
Model:	OLS	**Adj. R-squared:**	0.251
Method:	Least Squares	**F-statistic:**	3859.
Date:	Sat, 16 Dec 2017	**Prob (F-statistic):**	0.00
Time:	22:31:34	**Log-Likelihood:**	-40216.
No. Observations:	34588	**AIC:**	8.044e+04
Df Residuals:	34584	**BIC:**	8.047e+04

The r-squared of the regression is only 0.251.

4 CONCLUSIONS AND DISCUSSION

In a word, we have performed exploratory analysis and linear regression with the electric power consumption as well as the temperature during the same time interval. With the average daily consumption analysis by four quarters, the energy consumption pattern seems to follow the similar trend in every 24 hours, while the temperature in the different seasons have critical influence on it. The linear regression with temperature and different time span within a day is conducted with data, however the r-square is 0.251. The result of regression presents that the equation does not fit the data well. There are some probably reasons:

- The temperature data and consumption data are from different data sources. As the consumption data is set as household in France, the temperature one is record in Paris.
- The regression model does not fit the data well. Compared to the multiple regression model we have conducted, the model mentioned represent the highest r-square even though less than 0.3.

As for future work, the model still has potential to improve, one with the more reasonable variables involved instead of the various models under limited time. On the other hand, the segmentation of data could be tried to display if there are fitted curve exist.

REFERENCES

[1] UCI Machine Learning Repository
http://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption

[2] MeteoBlue website
https://www.meteoblue.com/en/weather/archive/export/paris_france_2988507

[3] Time and Data website
https://www.timeanddate.com/holidays/france/2010