Project Report
: Analysis on Energy Consumption of Appliances in a Low-Energy House

1. Abstract

This study's objective is to understand the energy consumption of appliances in a certified low energy house. Considering various predictors such as room temperature, humidity, the weather outside, and time, we performed a linear regression model to analyze the relationship of these factors with energy consumption. The best result obtained from the linear regression model was a R^2 score of 0.2881, with 25 variables considered significant. According to this model, predictors such as room 2 temperature, room 3 humidity, and outside humidity are important factors that influence energy consumption. However, examining the model diagnostics leads to the conclusion that regression analysis may not be the best model for this dataset, considering the Ljung-Box test or the ACF values. Furthermore, by applying spectral analysis, we discovered frequency cycles within the given timeline, and patterns of energy consumption in the house. Based on this analysis, we have concluded that the energy consumption follows a similar weekly pattern, (increasing on weekends and decreasing on weekdays)

2. Introduction

a. Analysis Objective

The main objective for this study is to analyze which factors contribute to how much of the energy consumption in the house and predict future consumption.

b. Motivation

i. Understanding which room contributes the most to the overall energy consumption is an important step to optimize energy usage in everyday life. By answering this question, we can diagnose ourselves in which rooms to be more aware of using electricity and replace appliances with more energy efficient ones.

ii. Also, understanding how the weather affects energy usage can provide the household useful information when planning monthly budgets. Being able to anticipate which points in the week or month would lead to more energy cost can guide the residents on how to allocate their spendings.

iii. According to other research, electricity demand showed a high dependence on the temperature of the days(1), and the daily energy usage patterns of the residents also was one of the strong factors that shaped energy consumption(2).

c. Data Explanation

This dataset is a collection of energy consumption(Wh) of a low energy house from 1/11/2016 5:00pm to 5/27/2016 6:00pm. Measurement took place every 10 minutes, resulting in 19,735 observations. There are also 29 input variables. The input variables include temperature(°C) and humidity(%) from eight rooms of the house, and the weather conditions such as wind speed(m/s) and pressure(mm Hg) that are measured at a nearby weather station.

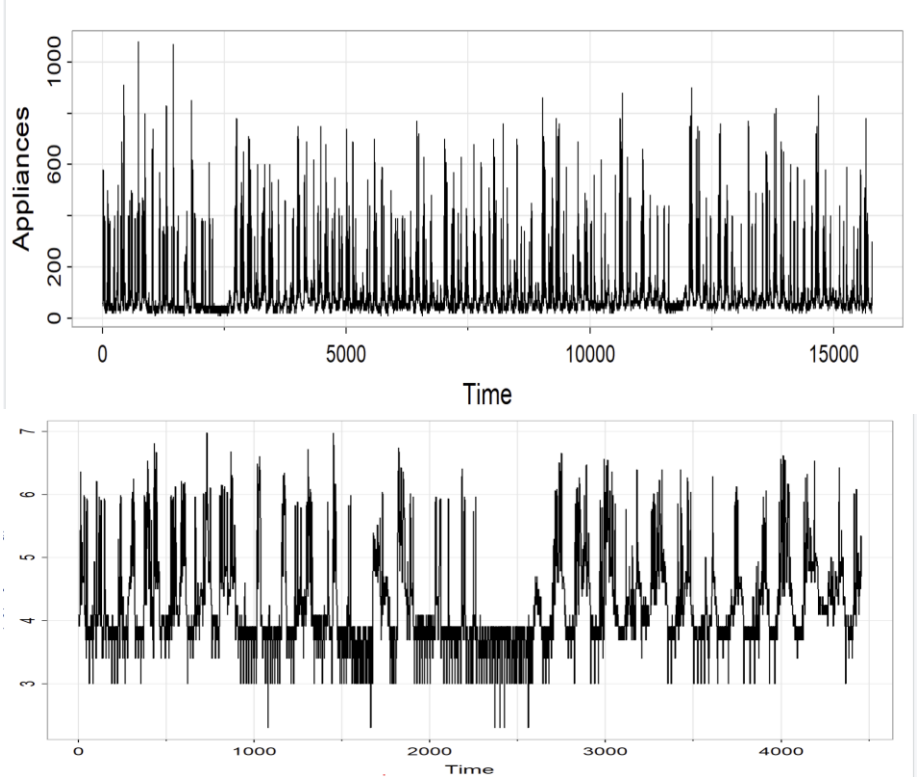
1)Hor, CL. "Analyzing the impact of weather variables on monthly electricity demand." *IEEE Trans. Power Syst.*, vol. 20, no. 4, 2005, pp. 2078-2085.

2)Saldanha, N. "Measured end-use electric load profiles for 12 Canadian houses at high temporal resolution." *Energy Build.*, vol. 49, no. -, 2012, pp. 519-530

3. Methods

a. Linear Regression

i. Preliminary Analysis



Based on the time series plot, we can see there isn't a particular trend of energy consumption, and the variance generally does not change throughout the period. The data does not seem stationary since the variance decreases as time passes. Therefore we applied log transformation on the output variable. The first plot above shows the plot of the whole dataset from 1/11/2016 5:00pm to 5/27/2016 6:00pm, while the lower plot gives us a closer view on just the January data that has been log-transformed.

ii. Regression Model I & II

We have performed a linear regression model based on the first 80% of the observations. By training the model this way we will test this data on the remaining 20%, and examine the prediction power through MSE. There were no missing values in the original dataset that we had to impute into other values, and removing outliers would not be appropriate considering the fact that energy consumption varies greatly. First, in order to perceive a general outline of the dataset, we have included all input variables to examine which ones are significant.

For the second regression model we have tried different approaches to improve the model. Since the dataset was not stationary, we have log transformed the output variable appliances. Also time did

not seem to have a linear combination with the output, therefore we added a quadratic term to it. Finally we excluded the insignificant variables that had a high p-value to simplify the model.

b. Spectral Analysis

i. Fast Fourier Transform

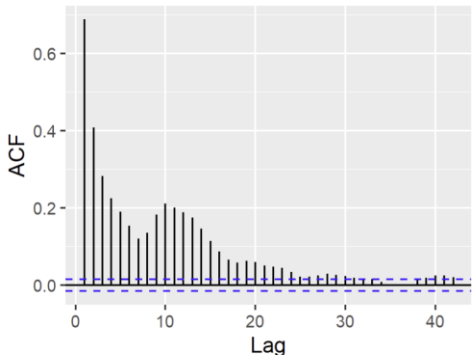
We will first use Fast Fourier Transform (FFT) the Appliances signal to get the magnitudes of the frequencies that construct the appliance signal. Then we are going to filter the frequencies with high magnitude and reconstruct the Appliance signal with those frequencies. We will analyze the spectral effect of each of the signals.

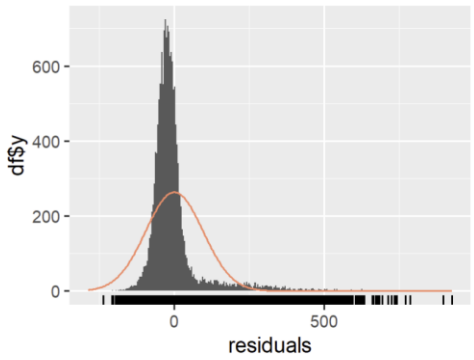
4. Results

a. Linear Regression

i. Regression Model I Results (All 29 variables included)

When including all the input variables, the linear regression returns a R^2 score of 0.1745. Similar to other previous studies, the model does not provide a strong explanation. This may be due to the fact that the Ljung -Box test rejects the null hypothesis rejecting that the auto-correlation values are zero. Also the ACF plot shows that the residuals are not completely white noise and that there may be some more information left in them.






```
Residual standard error: 0.5722 on 15761 degrees of freedom
Multiple R-squared:  0.2863,    Adjusted R-squared:  0.2851
F-statistic: 243.1 on 26 and 15761 DF,  p-value: < 2.2e-16
```

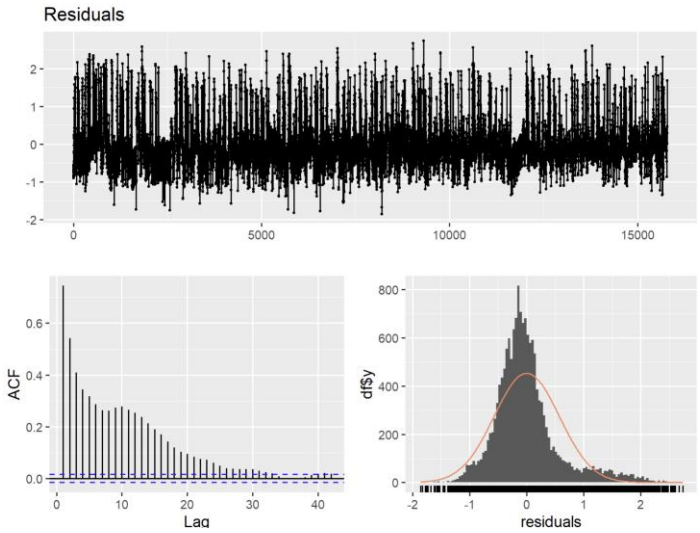
```
        Ljung-Box test

data:  Residuals
Q* = 14836, df = 10, p-value < 2.2e-16
```

ii. Regression Model II Results

The second model resulted a R^2 score of 0.2871. Even though the output was log-transformed and the insignificant input variables were removed, the overall result did not change meaningfully. The Ljung Box test still

showed that the auto correlations are not zero, and likewise, the ACF plot indicates that the residuals are not truly white noise.

```
Residual standard error: 0.5714 on 15763 degrees of freedom
Multiple R-squared:  0.2882,    Adjusted R-squared:  0.2871
F-statistic:   266 on 24 and 15763 DF,  p-value: < 2.2e-16
```



iii. Final Model Selection

Comparing AIC, BIC would be the most appropriate way to compare the two models, since their MSE would be different due to the log transformation in the second model. Comparing the AIC, BIC, R^2, gives us the conclusion the second regression model performed a better explanation. The second model's AIC, BIC were lower while its $R^2 (0.2882 > 0.1745)$ was higher.

```
> AIC(reg1)
[1] 188845.2
> AIC(reg2)
[1] 27158.14
> BIC(reg1)
[1] 189059.9
> BIC(reg2)
[1] 27357.48
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.732e+00  1.511e-01  24.704  < 2e-16 ***
time(date)    -6.550e-05  7.090e-06  -9.237  < 2e-16 ***
I(time(date)^2) 2.205e-09  3.244e-10   6.796 1.12e-11 ***
lights         1.725e-02  6.254e-04  27.578  < 2e-16 ***
T1            -3.072e-02  1.342e-02  -2.289 0.022090 *
RH_1           9.182e-02  4.493e-03  20.434  < 2e-16 ***
T2            -8.015e-02  1.250e-02  -6.411 1.48e-10 ***
RH_2          -7.396e-02  5.368e-03 -13.778  < 2e-16 ***
T3             1.824e-01  7.643e-03  23.865  < 2e-16 ***
RH_3           3.619e-02  5.129e-03   7.056 1.79e-12 ***
T4            -2.200e-02  6.458e-03  -3.407 0.000658 ***
RH_4          -2.726e-02  4.534e-03  -6.012 1.87e-09 ***
RH_5           2.616e-03  5.790e-04   4.518 6.28e-06 ***
T6             5.222e-02  4.992e-03  10.461  < 2e-16 ***
RH_6           1.781e-03  5.382e-04   3.309 0.000939 ***
T7            -5.253e-02  9.346e-03  -5.621 1.93e-08 ***
RH_7           7.174e-03  2.919e-03   2.458 0.013990 *
T8             1.288e-01  6.909e-03  18.644  < 2e-16 ***
RH_8          -5.976e-02  2.623e-03 -22.781  < 2e-16 ***
T9            -6.991e-02  1.240e-02  -5.640 1.73e-08 ***
RH_9          -1.294e-02  2.899e-03  -4.462 8.16e-06 ***
T_out         -1.515e-02  5.878e-03  -2.577 0.009986 **
RH_out         4.498e-03  8.985e-04   5.006 5.62e-07 ***
Windspeed      9.835e-03  2.246e-03   4.379 1.20e-05 ***
Visibility     9.599e-04  3.751e-04   2.559 0.010505 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5714 on 15763 degrees of freedom
Multiple R-squared:  0.2882,    Adjusted R-squared:  0.2871
F-statistic:   266 on 24 and 15763 DF,  p-value: < 2.2e-16
```

iv. Forecast of Future Five Values

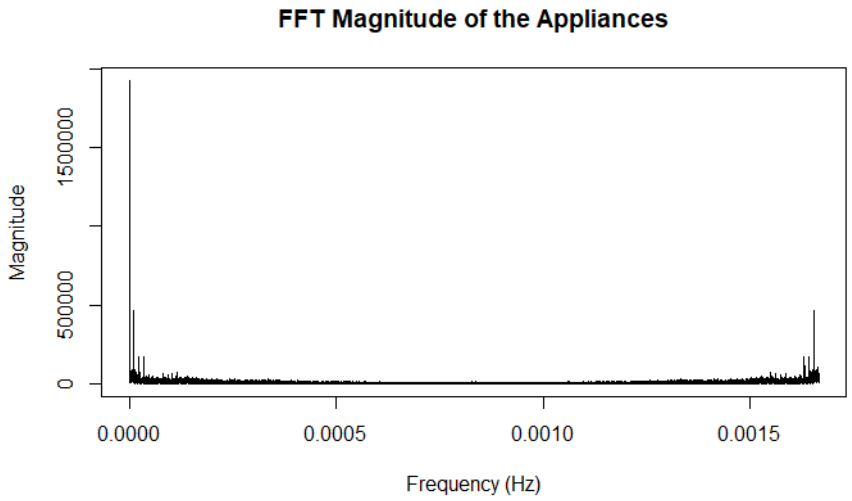
With the predict() function, we can forecast the future five values of the data and compare it with the actual next five values. Considering the fact that the output variable is

log transformed, we have also log transformed the actual five values to compare.

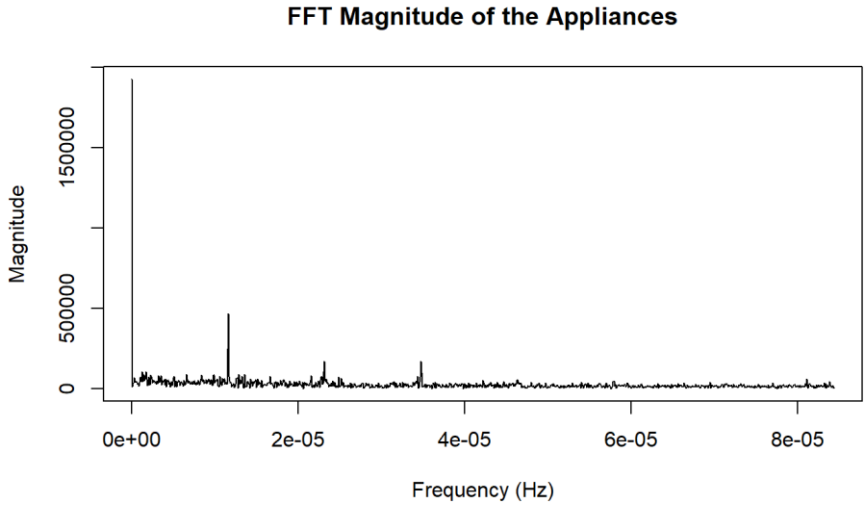
```
> predict(reg2, df[15788:15793,])
   15788    15789    15790    15791    15792    15793
4.554286 4.543875 4.558177 4.564981 4.748895 4.613458
> log(df[15788:15793,2])
[1] 5.703782 5.913503 6.380123 5.768321 5.736572 5.560682
```

b. Spectral Analysis

i. FFT Result

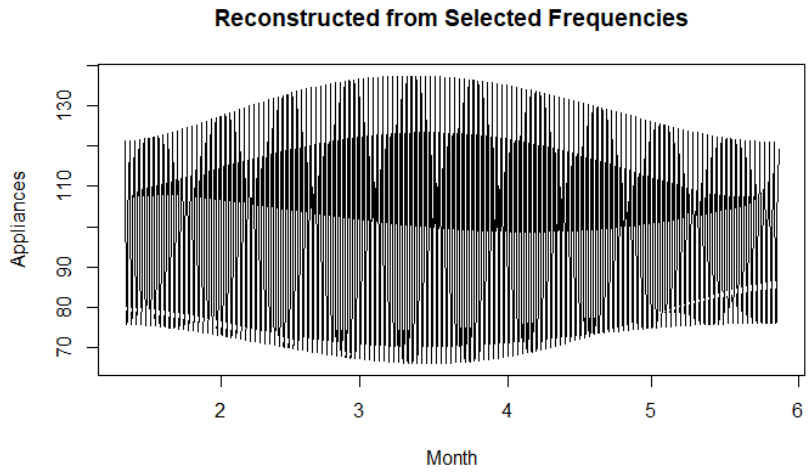


This plot shows the frequencies. Sampling rate is set to 6/3600 as the data time stamp has a 10 minutes interval so there are 6 data in an hour. We can see high magnitudes in the lower frequency area. To filter only the frequencies with high magnitudes, we expanded the image and checked the magnitudes.

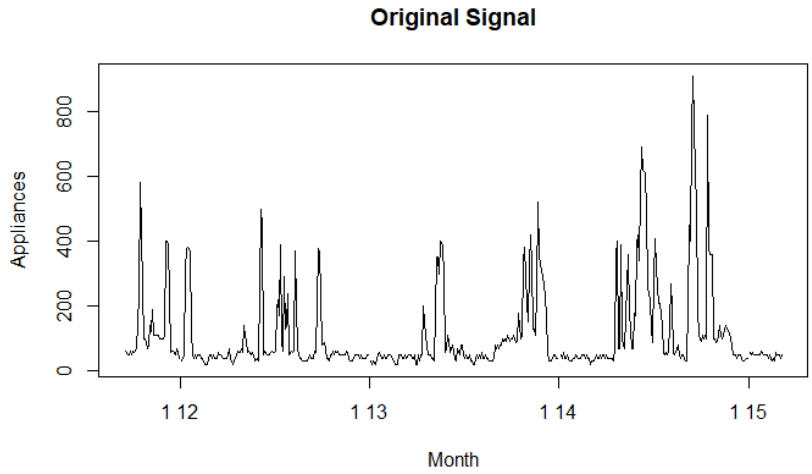


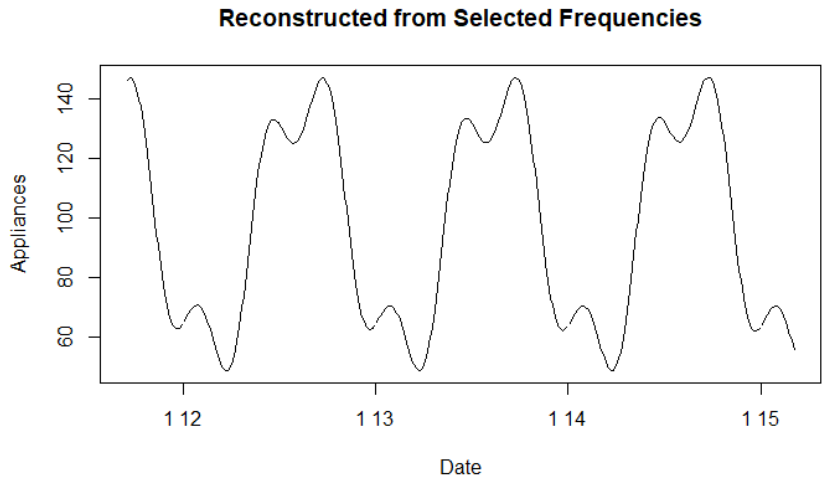
There are four spikes of frequencies with high magnitude. We will use these frequencies to reconstruct the appliances signal.

ii. Reconstruction of Filtered Frequencies



We filtered the four predominant signals and reconstructed the Appliance signal. This plot shows the signal of the whole period. For closer examination, we checked the date from 1/12/2016 to 1/15/2016.



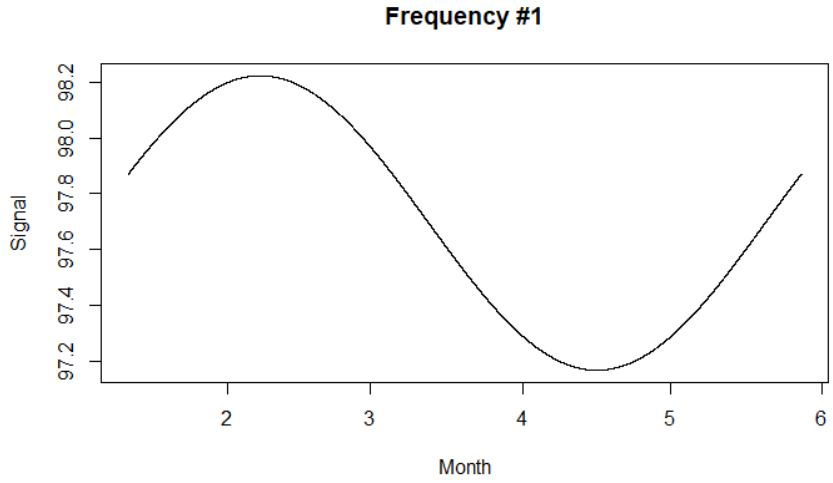


From the original signal we filtered the predominant frequencies and reconstructed a periodic graph.

frequencies <dbl>	magnitudes <dbl>
0.000000e+00	1928010.0
1.156997e-05	466585.5
3.470991e-05	171567.9
2.313994e-05	170487.8

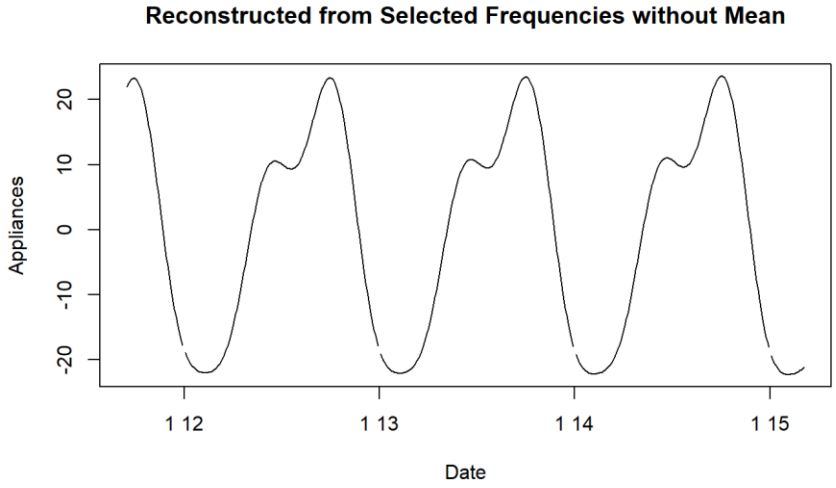
The values on the table show the top 4 magnitudes and the frequency values. We examined the four frequencies one by one.

1. Frequency #1



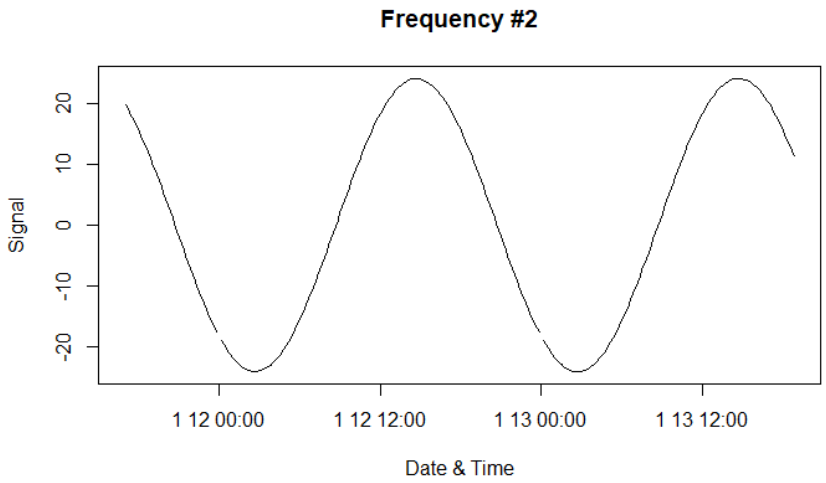
This plot shows the signal of the biggest magnitude. It shows periodicity of about 6 months. And the amplitude is about 97.2 to 98.2.

This is from 0 Hz frequency. It represents an overall mean of the data.



If we filter the mean component out, we get Appliances signal with mean around 0. We want to see the actual value of Appliances, so we considered the mean component in our analysis.

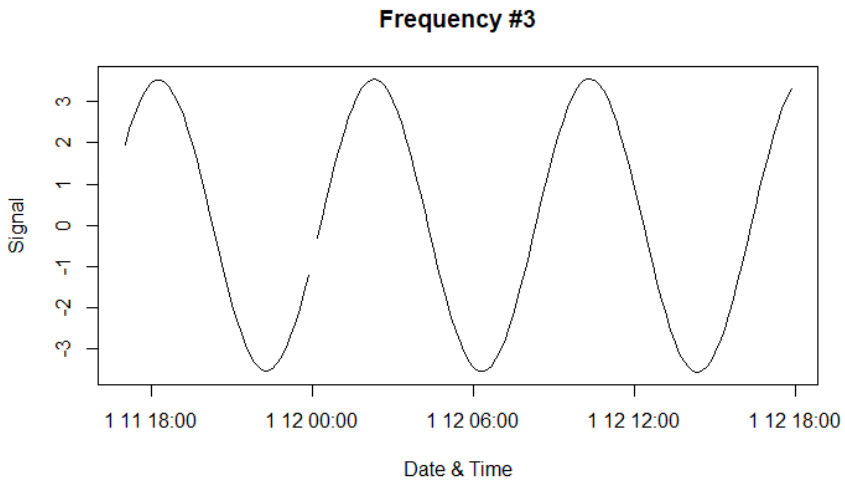
2. Frequency #2



This plot shows the signal of the second biggest magnitude. It shows periodicity of about one day. And the amplitude is about -20 to 20. Except for the mean frequency #2 is the most dominant component.

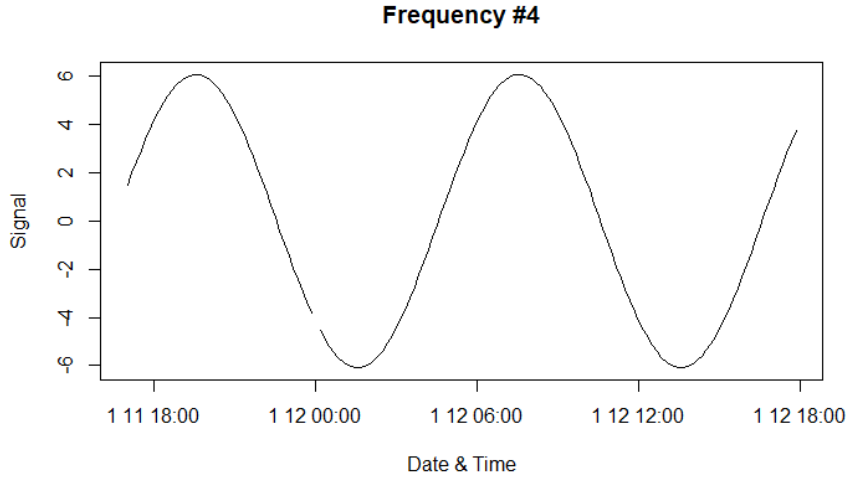
The magnitude of frequency #2 (466585.5) is about 2.7 times bigger than the following frequency (171567.9).

3. Frequency #3



This plot shows the signal of the third biggest magnitude. It shows periodicity of about 8 hours. And the amplitude is about -20 to 20.

4. Frequency #4



This plot shows the signal of the fourth biggest magnitude. It shows periodicity of about 12 hours. The amplitude is about -6 to 6.

To wrap up the four frequencies results, we can conclude that the cycle of the predominant components align with our expectation. The mean value (Frequency #1) goes up during winter and summer season and these are the expected periods when energy is used a lot. The second dominant component (Frequency #2) shows a daily cycle which also aligns with common expectations of energy usage.

5. Discussion

Examining the linear regression results have shown us that the linear combination of the input variables does not explain the energy consumption variance significantly. With 0.2882 as the highest R^2, the two models' ACF plot and Ljung-Box test indicate that the residuals were not close to white noise. This means that the model was not able to capture the variance of the energy consumption properly.

Returning to the motivation and objective of this study, when trying to discover which rooms or weather conditions are important we can refer to the coefficients and the p-values of the input variables from model 2. Conditions such as room 2 humidity, room 3 temperature, humidity outside can be classified as important based on this model, however we must keep in mind that there may be other explanations that the model does not provide us. For further research to answer the motivation and goal for this study, we can apply the regression model to the refined data acquired by fast fourier transform or use other modeling techniques such as random forest or SVM, that may capture the information that were left in the residuals in the regression model.

6. Team Contribution

a. [redacted] : Linear regression, descriptive analysis
b. Hyomin Yoo : Spectral Analysis

7. Appendix

a. Data source
: UCI Machine Learning Repository, Appliances Energy Prediction
https://archive.ics.uci.edu/dataset/374/appliances+energy+prediction

b. Data dictionary

Variables	Description
date time	Year-month-day hour:minute:second
Appliances	Energy use in Wh
lights	Energy use of light fixtures in the house in Wh
T1	Temperature in kitchen area, in Celsius
RH_1	Humidity in kitchen area, in %
T2	Temperature in living room area, in Celsius
RH_2	Humidity in living room area, in %
T3	Temperature in laundry room area
RH_3	Humidity in laundry room area, in %
T4	Temperature in office room, in Celsius
RH_4	Humidity in office room, in %
T5	Temperature in bathroom, in Celsius
RH_5	Humidity in bathroom, in %
T6	Temperature outside the building (north side), in Celsius
RH_6	Humidity outside the building (north side), in %
T7	Temperature in ironing room, in Celsius
RH_7	Humidity in ironing room, in %
T8	Temperature in teenager room 2, in Celsius
RH_8	Humidity in teenager room 2, in %
T9	Temperature in parents room, in Celsius
RH_9	Humidity in parents room, in %
To	Temperature outside (from Chièvres weather station), in Celsius
Pressure	Pressure (from Chièvres weather station), in mm Hg
RH_out	Humidity outside (from Chièvres weather station), in %
Windspeed	Windspeed (from Chièvres weather station), in m/s
Visibility	Visibility (from Chièvres weather station), in km
Tdewpoint	Tdewpoint (from Chièvres weather station), °C
rv1	Random variable 1, nondimensional
rv2	Random variable 2, nondimensional