Richard Swann | 823902440
Jason Cao | 200107583
Matt Fung | 301329703

CMPT318 Assignment 2

Q1.

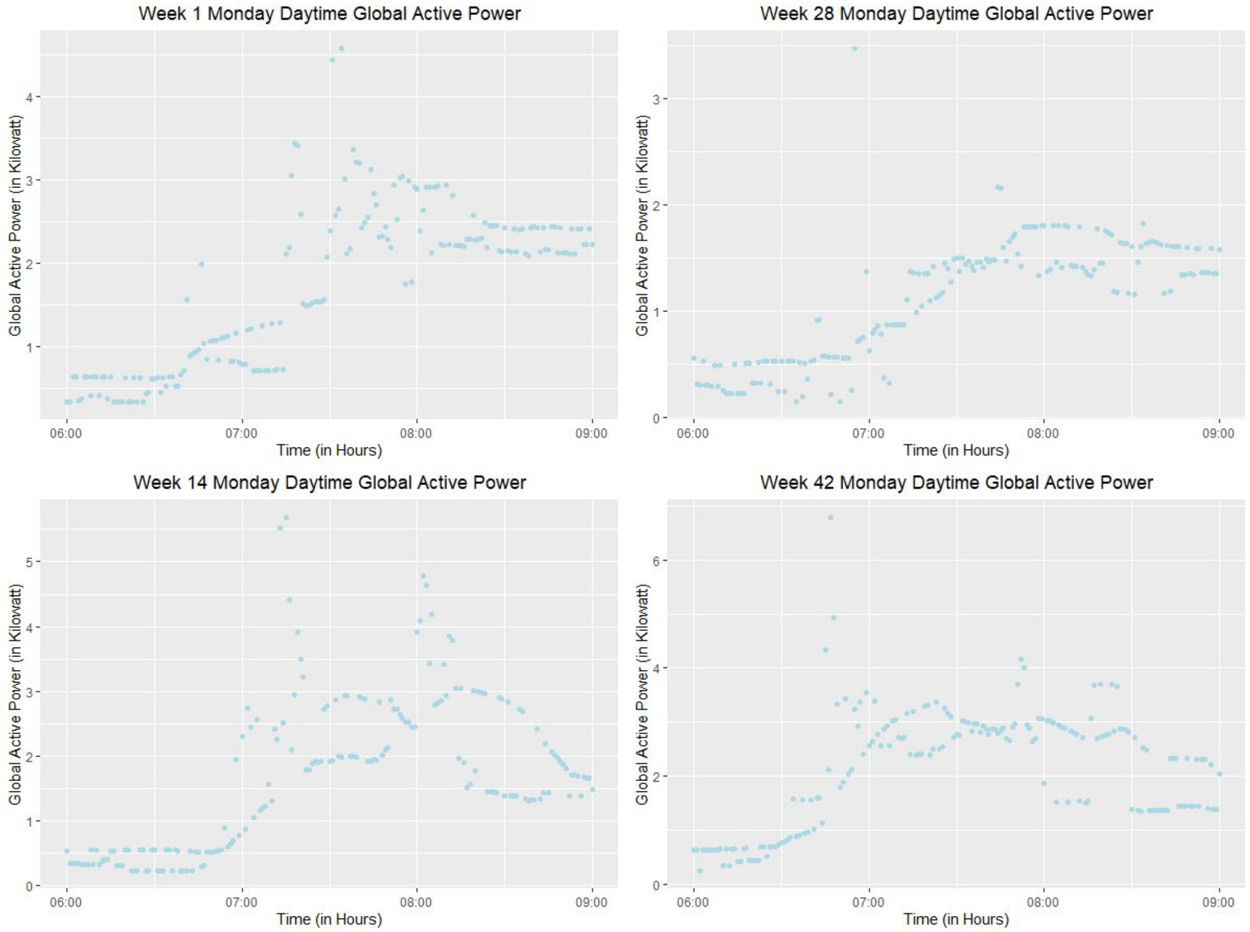


Figure 1. Diagram Depicting the Plots of Global Active Power for Mondays Daytime For Weeks 1, 14, 28 and 42

Comments:

We decided to determine a time window on weekdays that shows a clearly recognizable electricity consumption pattern over a time period of several hours for Global_active_power. From Figure 1 above, we determined our time window to be between 6:00 AM - 9:00 AM for Mondays. We can observe that there is a surge in Global Active Power between 6:00 AM - 9:00 AM where the magnitude of Global Active Power peaks between 7:00 AM and 8:00 AM for weeks 1, 14, 28, 42. The reasoning behind why this pattern occurs can be explained by the nature of a typical person's daily routine. The typical person works a daytime job with work hours spanning between sometime in the morning to sometime in the late afternoon/evening. Hence, the typical person will most likely wake up sometime between 6:00 - 9:00. Once awoke, the typical person would likely have to get ready for work. Although each person's morning routine is different, it is extremely likely that electricity is used. A surge in Global Active Power can be interpreted as people using electricity for their morning routine activities. Such examples include taking a shower, using a blowdryer, turning on the lights, using an oven, turning on the TV for morning news etc. Once the typical person has completed their morning routine, they will most likely leave for work. As a result of this, no more electricity is used in their homes. Thus, after the surge in Global Active Power between 6:00 AM - 9:00 AM, the magnitude of Global Active Power begins to gradually decrease in the next few hours for weeks 14, 28, and 42. (With the exception of week 1)

Q2.

A series of 2.5-hour time windows (Monday 7:00 AM to 9:30 AM) of Global Intensity measurements were used to generate 20 HMM models with each model consists a particular number of states.

Aside from fitting the data to a model with a specified state, depmix() also helps to generate corresponding BIC and log-likelihood for each model. These value pairs are used as the criterion for selecting the “optimal” HMM model.

States	BICs	Log-Likelihood
0	0.000	0.0000
2	38192.253	-19064.7365
3	25956.175	-12915.3079
4	24079.236	-11936.4802
5	23279.996	-11487.5329
6	20170.815	-9874.6470
7	19417.757	-9430.8542
8	7280.362	-3285.9245
9	5457.674	-2289.3791
10	-16689.761	8878.5075
11	6839.217	-2782.8434
12	6651.832	-2577.0444
13	-1277.139	1508.5162
14	-16240.576	9120.2783
15	3764.245	-743.1202
16	-29939.563	16256.7646
17	-43177.357	23032.6109
18	-38241.805	20730.7525
19	-74072.869	38821.1704
20	-119952.841	61945.0113

Table 1. Values of BICS and Log-Likelihood For N-Number of States For Dataset 3.

The “optimal” HMM model is the one with a negative log-likelihood (-743.1202) closest to 0 and a relatively high BIC (3764.245). This conclusion is verified by the plot below. Model with 15 states seems to be the best fit HMM.

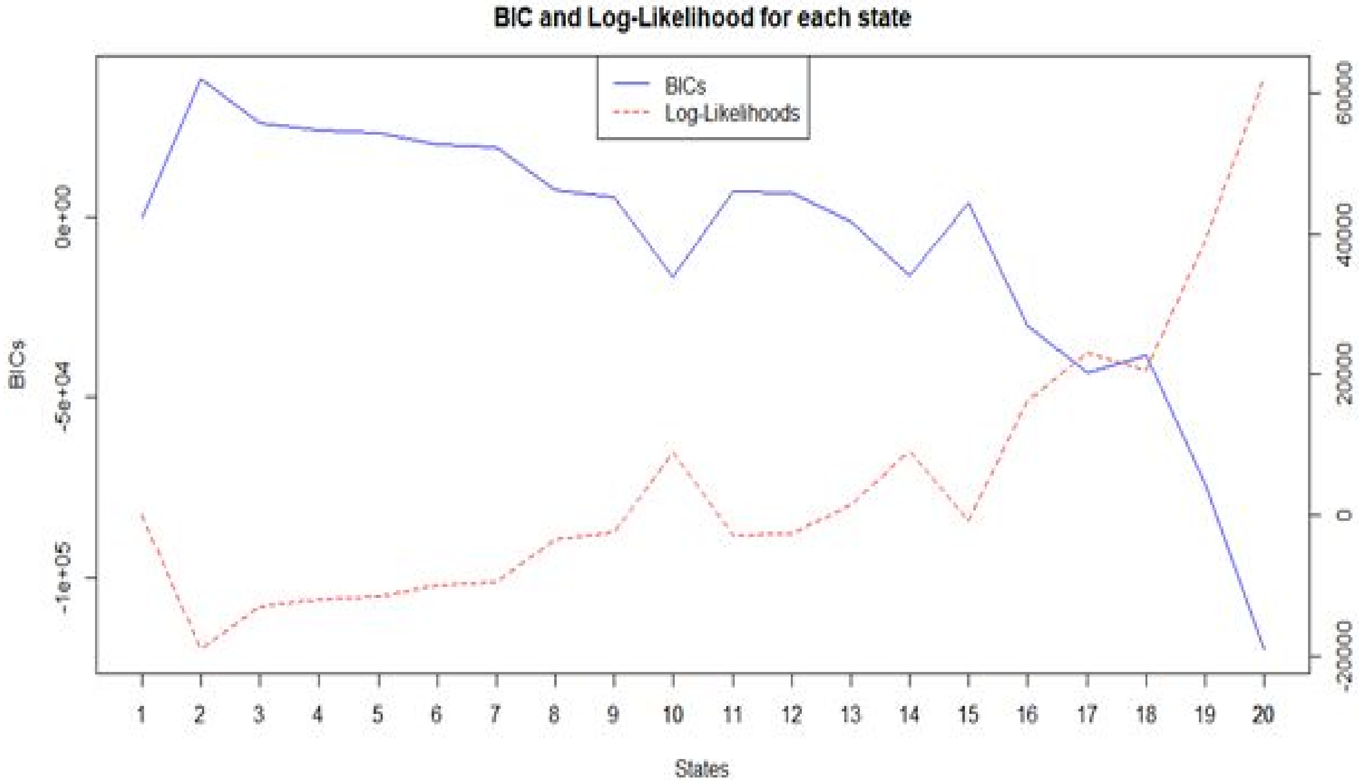


Figure 2. Plot of BIC and Log-Likelihood For Each State

N.B Warnings were generated in the process of running depmix(). (will ask how to explain warnings during office hours.)

Q3.

This question investigates using deviation from a “moving average” as an anomaly detection method.

The data for week 11 was arbitrarily chosen for evaluation. After inspecting the plots of “Global_active_power”, “Global_reactive_power” and “Global_intensity”, the predictor “Global_intensity” was selected for further study as it appeared to be the most regular of the 3 predictors.

The Week 11 Global Intensity vector extracted from the main dataset had 10080 values with a maximum value of 37.8.

An Anomaly Detection Threshold (ADT) was selected, again somewhat arbitrarily, at 20% of the maximum value of the Global Intensity predictor. The following graph shows the data, the moving average and the positive and negative thresholds.

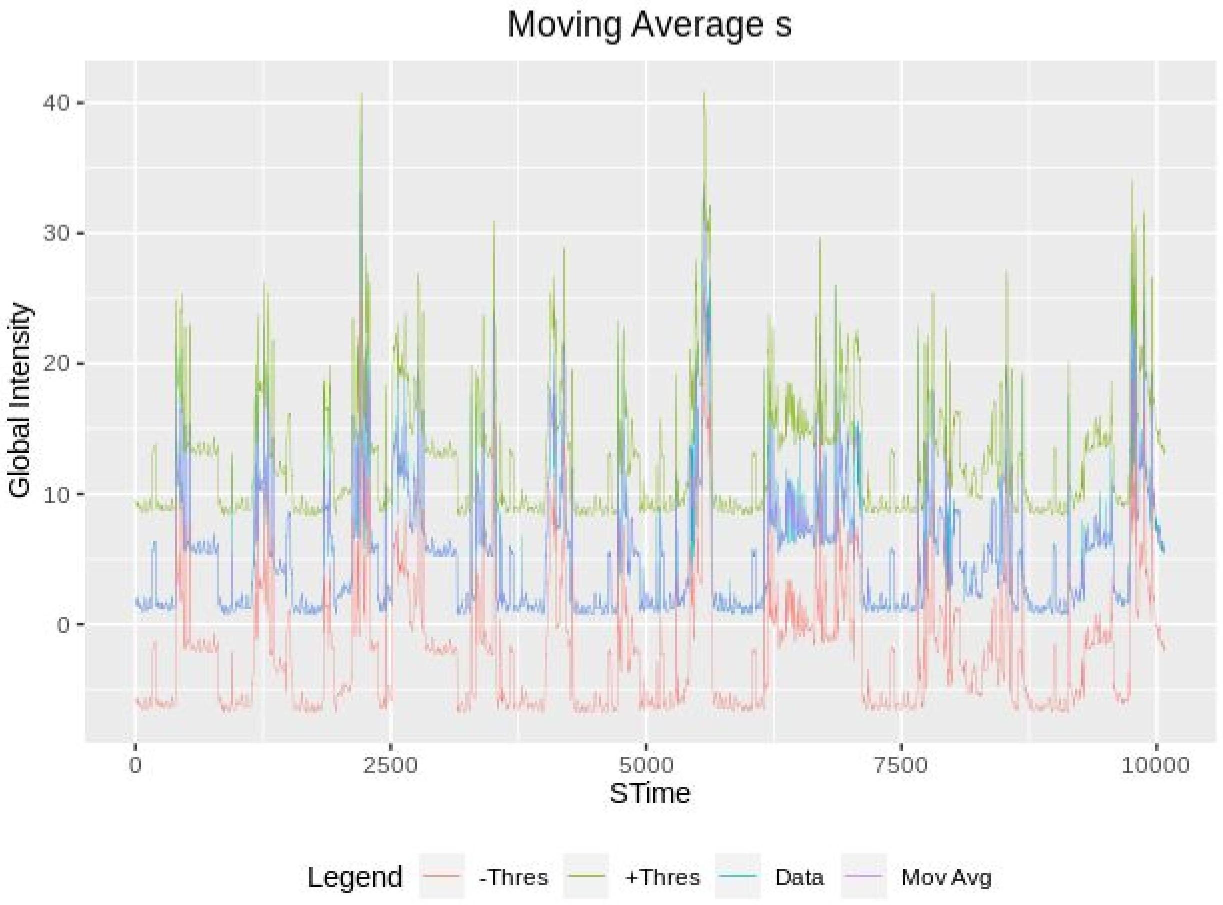


Figure 3. Plot of the Moving Average and the Positive and Negative Thresholds.

The “Practical Math” R library “pracma” was loaded and the moving average function “movavg” was used to evaluate the performance of a moving average for anomaly detection.

Various moving average window lengths from 2 to 20 were evaluated for the number and magnitude of the differences exceeding the ADT. Not surprisingly, smaller windows, reduced the number of anomalies detected and longer windows increased them. A window length of 10 was selected as a reasonable compromise between predictive ability and excessive false positives.

The “movavg” function provides a variety of methods for calculating a moving average. The decision was made to evaluate these different methods to see if there was a significant difference in their performance.

The calculations supported are:

- s for ``simple'', it computes the simple moving average.
- t for ``triangular'', it computes the triangular moving average by calculating the first simple moving average with window width of ceil(n+1)/2; then it calculates a second simple moving average on the first moving average with the same window size.
- w for ``weighted'', it calculates the weighted moving average by supplying weights for each element in the moving window. Here the reduction of weights follows a linear trend.
- m for ``modified'', it calculates the modified moving average. The first modified moving average is calculated like a simple moving average. Subsequent values are calculated by adding the new value and subtracting the last average from the resulting sum.
- e for``exponential'', it computes the exponentially weighted moving average. The exponential moving average is a weighted moving average that reduces influences by applying more weight to recent data points () reduction factor 2/(n+1);
- r for``running'', this is an exponential moving average with a reduction factor of 1/n.

It was decided to evaluate “s”, “t”, “w” and “e” calculations. The results were significantly different and are summarized in Table Q3-1, below. “Maximum difference” was calculated as the maximum of the absolute value of the raw data point minus the equivalent weighted average point. The “Anomaly count” is the number of differences (as previously calculated) that exceed the Threshold (20% of the maximum Global Intensity reading).

Calculation	**s**	**t**	**w**	**e**
Maximum difference	15.64	15.48	12.61	13.52
Anomaly count	144	132	56	53

Table 2. Performance of different moving average calculations

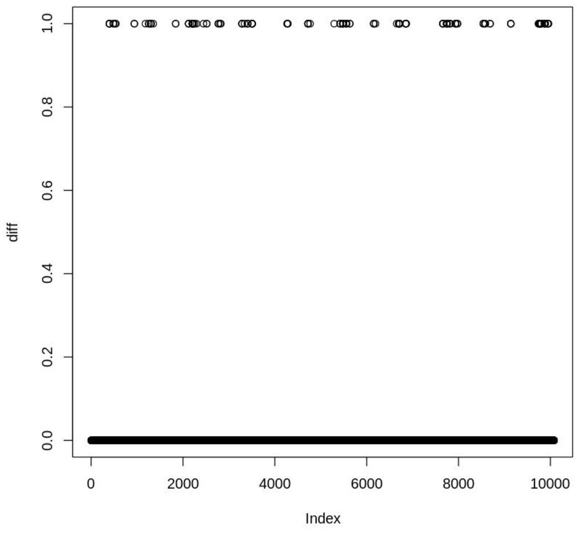
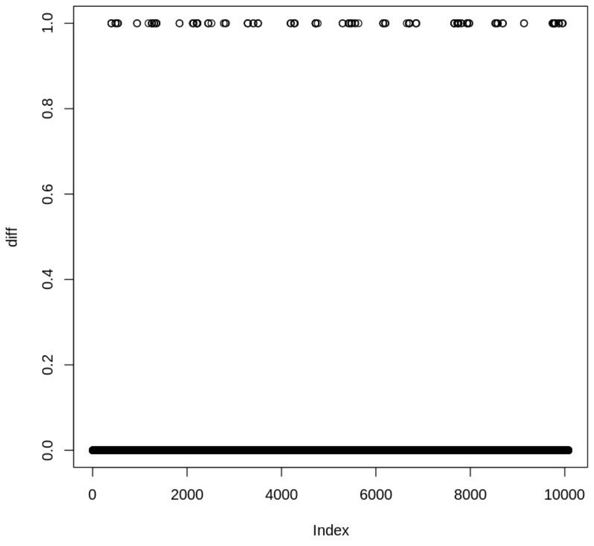
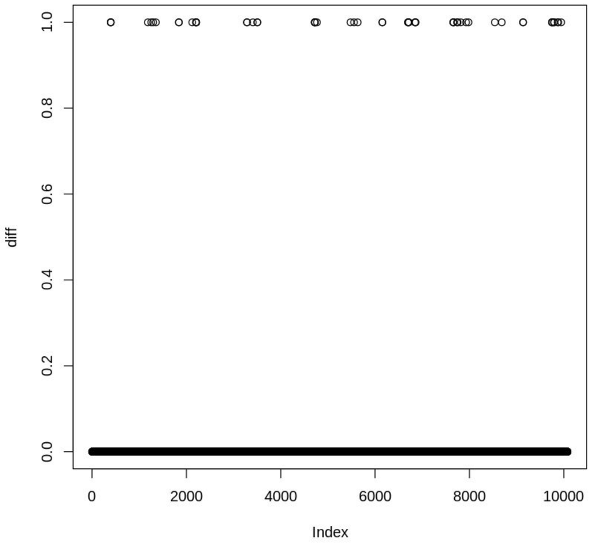
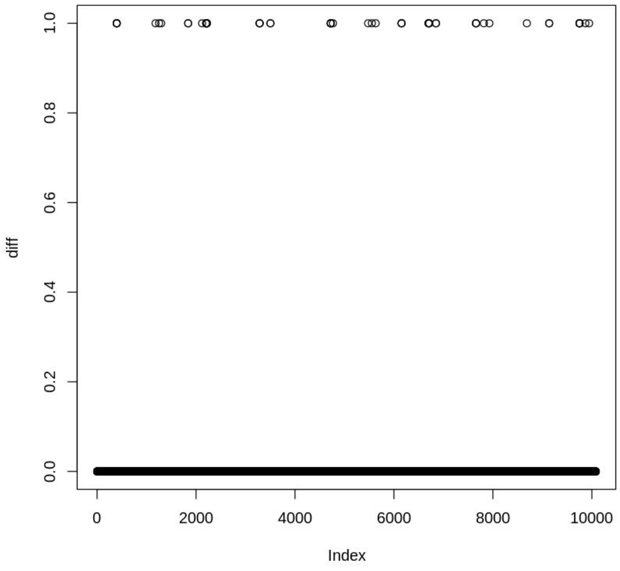
Calculation “s”	Calculation ‘t’
	
Calculation “w”	Calculation “e”
	

Table 3. Distribution of “Anomalies” based on calculation type

Conclusions

Using a threshold of 20% of the maximum value, a moving average window of 10 observations and a “weighted moving average” calculation suspected anomalies might be detected with a false positive rate of 53 per 10080 observations or 8 per day. It is possible that this approach could be refined to be a useful anomaly detection tool.