CMPT 318
Term Project

Spring 2020

Group 12

Yoonhong Lee 301267876
Chester Cervantes 301291560
Yevhenii Strilets 301808011
Jimmy Tran 301218618
Ilia Krasavin 301326046

Table of Contents

List of Tables

List of Figures

Abstract

Electricity power grids frequently rely on supervisory control and data acquisition, hence the project explores behaviour-based intrusion detection methods used for cyber situational awareness of automated control processes. Nowadays, SCADA control system architecture allows continuous analysis of real-time control data and detects potentially harmful anomalies [1]. However, during the data analyzing process, challenges such as imperfections in the data, corrupted values, various types of anomalies arise and disturb the process of creation of stochastic models and anomaly

detection. Understanding and analyzing the dataset of the electric power grid allows creating a suitable probabilistic model such as the Hidden Markov model.

Introduction

The project explores behaviour-based anomaly detection methods in terms of intrusion detection methods used for cyber situational awareness in the analysis of automated control processes. The dataset contains 4 years' worth of electricity consumption data with nine different features, which will be used in the creation of the Hidden Markov model for the analytical approach of anomaly detection for different types of simple and complex anomalies. The whole process is divided into four main steps. The first step is to explore the data and gain a better understanding of the basic data characteristics. Also, this step contains the identification of the specific window that displays recognizable electricity consumption patterns. The second step contains feature engineering to find a suitable set of dependent variables for Hidden Markov models that would optimize the accuracy and minimize the complexity of the models. Also, analyze the correlation between the features and implement the PCA in order to visualize how the data is spread out in a dataset. The third step represents the training and testing of the Hidden Markov Model and finding the optimal number of states to eliminate the overfitting and underfitting on the data. The last part contains the single week of data with injected anomalies and requires the application of the Moving Average method in order to receive a reasonable most recent observation and provide a

reasonable expectation for the next observation. All of the above steps give a clear understanding and suitable probabilistic model for the purpose of representing ‘normal’ system behaviour to the extent possible with no ground truth available.

Problem Scope

This project is for analyzing the 4-year electricity power dataset which is partitioned minute by minute. Analyze of everyday data will be used for determining the pattern of electricity consumption in terms of global active power, global intensity and voltage. The time window for weekdays and weekends will be stated according to analyzed data and its pattern. Based on retrieved correlation information from data features, the global active power will be used for training univariate Hidden Markov. While for multivariate Hidden Markov models the combination of global active power, global intensity power and voltage will be used. At the same time, there will be chosen an optimal number of states by comparing the bias and variance of models based on their log-likelihood and BIC values. The above steps will allow eliminating the overfitting and underfitting of data. The same time window will be used for anomaly detection in five different datasets, while the application of the Moving Average method will allow the detection of complex anomalies.

Methodology

This project is divided into four steps and each step contains a specific methodology. The descriptive statistics method will be used for summarization of data features from an electricity consumption dataset using indexes such as the mean, standard deviation, mode and geometrical mean. Also, the Population Pearson Correlation of each disjoint pair of features will be computed and represented with a correlation matrix to represent the most correlated pairs. Moreover, the data retrieved from the correlation matrix will be used to determine the dependent variable as well as a combination of several dependent variables for training univariate and multivariate Hidden Markov models on normal electricity consumption data. The random week of data will be retrieved from five anomaly injected datasets and analyzed with the Moving Average method.

Data Exploration

This section aims to find a time window that shows clear observable patterns in both weekdays and weekends. The dataset was analyzed and contained a large number of NA sections. Hence, the dataset has been filtered by removing all rows that contain NAs values for all response variables to avoid future complication in calculations in R. The purpose of the data exploration phase is getting a better understanding of the basic data characteristics and statistical knowledge on each response variable.

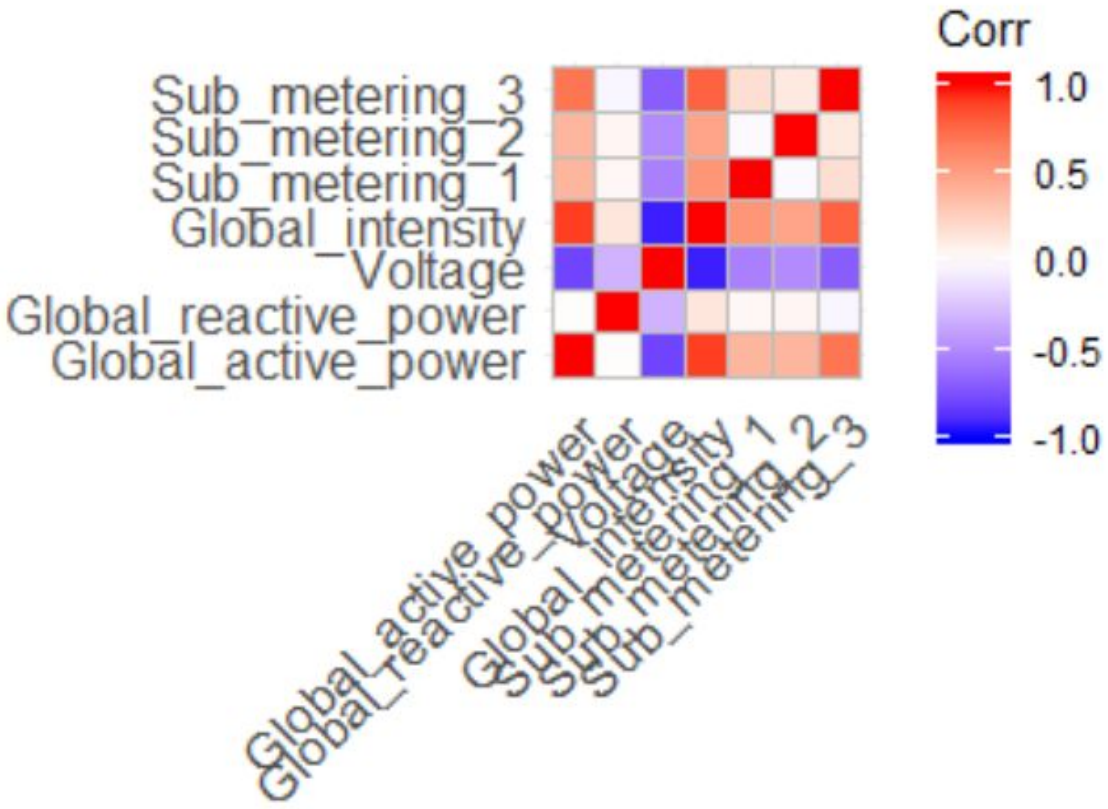
Therefore, arithmetic mean, geometric mean, median, mode, and standard deviation of all response variables is calculated.

DESCRIPTIVE STATISTICS METRICS

Data	Mean	Geom. mean	Median	Mode	SD
Global Active Power	1.227778	0.8580273	0.815825	0.218	1.057457
Global Reactive Power	0.1218387	0.0	0.1	0.0	0.1116499
Voltage	240.5478	240.5256	240.78	241.18	3.261038
Global intensity	4.64223523	2.9470089	2.6	1	0.1116522
Sub metering 1	1.15603166	0.0	0.0	0	6.2753504
Sub metering 2	1.35898807	0.0	0.0	0	6.0234654
Sub metering 3	6.1666531	0.0	0.0	0	8.3151377

Looking at table 1, it appears that all there are small differences between mean, geometric mean, and medians for all response variables. Also, the range of all standard deviations are quite small as well. This suggests that given data sets very well follow the law of large numbers and possibilities for errors in our data sets are very low [2]. Therefore, the dataset can be used in further data exploration and search for relationships between response variables. The table named Correlation of Disjoint Pairs displays the calculated correlation of all disjoint pairs of response variables and checks for any relationship that response variables make on other response variables.

CORRELATION OF DISJOINT PAIRS



Looking at the correlation matrix above, this does suggest that disjoint pairs of Global Active Power, Global Intensity, and Voltages show the best correlations. The specific summary of all the correlations are shown below in the table.

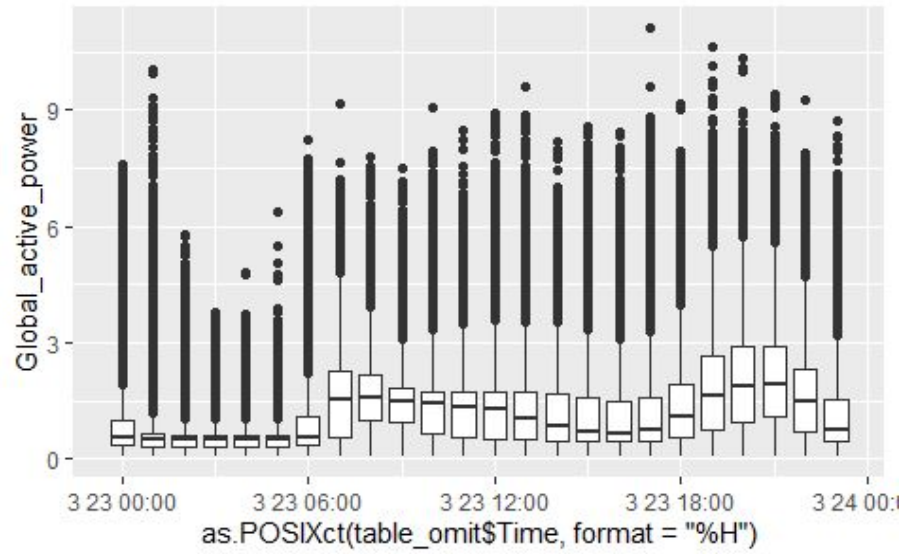
CORRELATION SUMMARY

High positive correlation	Global Intensity & Global Active Power
Moderate positive correlation	Global Active Power & Sub Metering 3 Global Intensity & Sub Metering 3
No correlation / Low Positive / Low Negative Correlation	Global Active Power & Global Reactive Power Global Active Power & Sub Metering 1 Global Active Power & Sub Metering 2 Global Reactive Power & Voltage Global Reactive Power & Global Intensity Global Reactive Power & Sub Metering 1

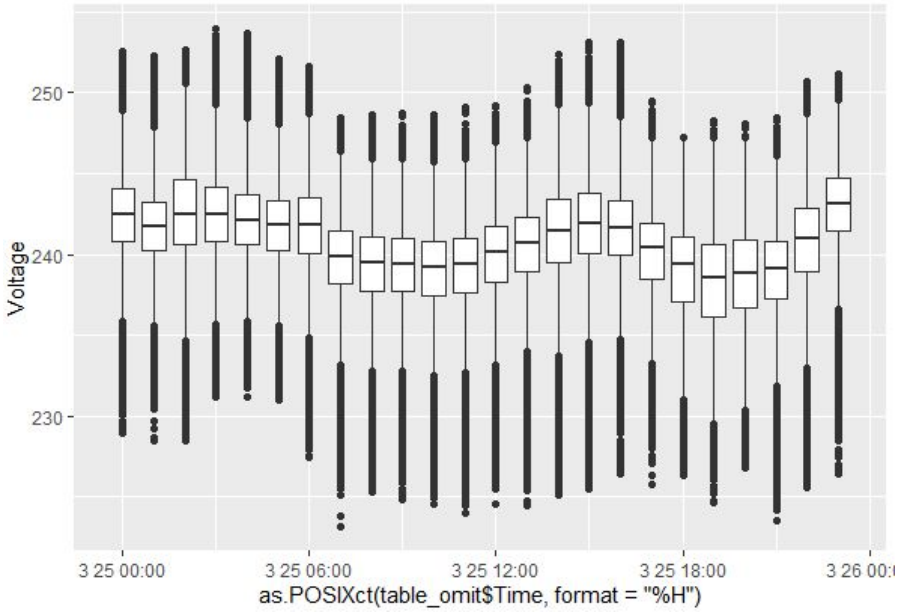
	Global Reactive Power & Sub Metering 2 Global Reactive Power & Sub Metering 3 Voltage & Sub Metering 1 Voltage & Sub Metering 2 Global Intensity & Sub Metering 1 Global Intensity & Sub Metering 2 Sub Metering 1 & Sub Metering 2 Sub Metering 1 & Sub Metering 3 Sub Metering 2 & Sub Metering 3
Moderate Negative Correlation	Voltage & Sub Metering 3
High Negative Correlation	Global Active Power & Voltage

Based on the correlation found above, Global Active Power, Global Intensity, and Voltage are plotted to confirm our correlation calculations.

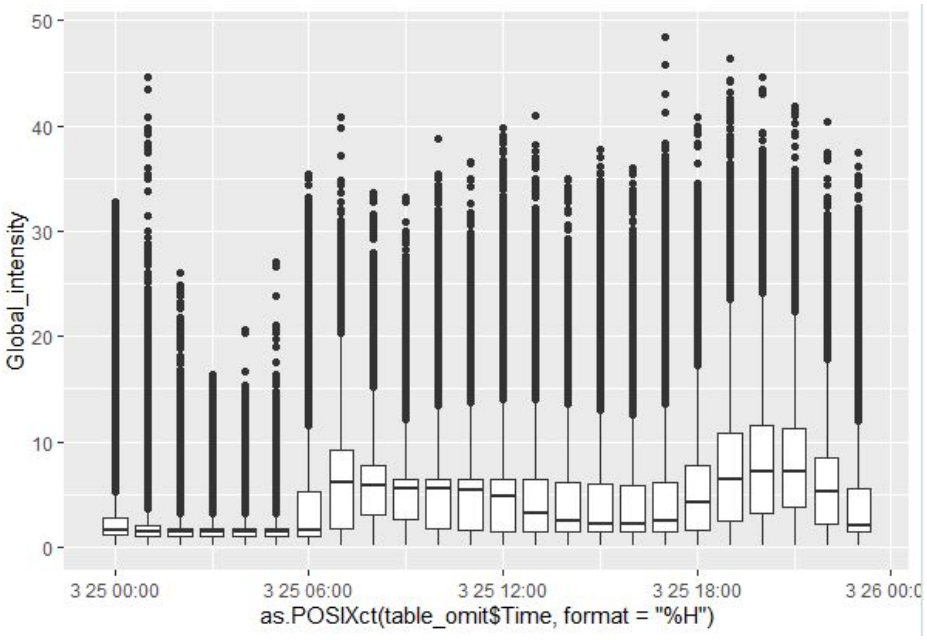
Global Active Power



Voltage



Global Intensity



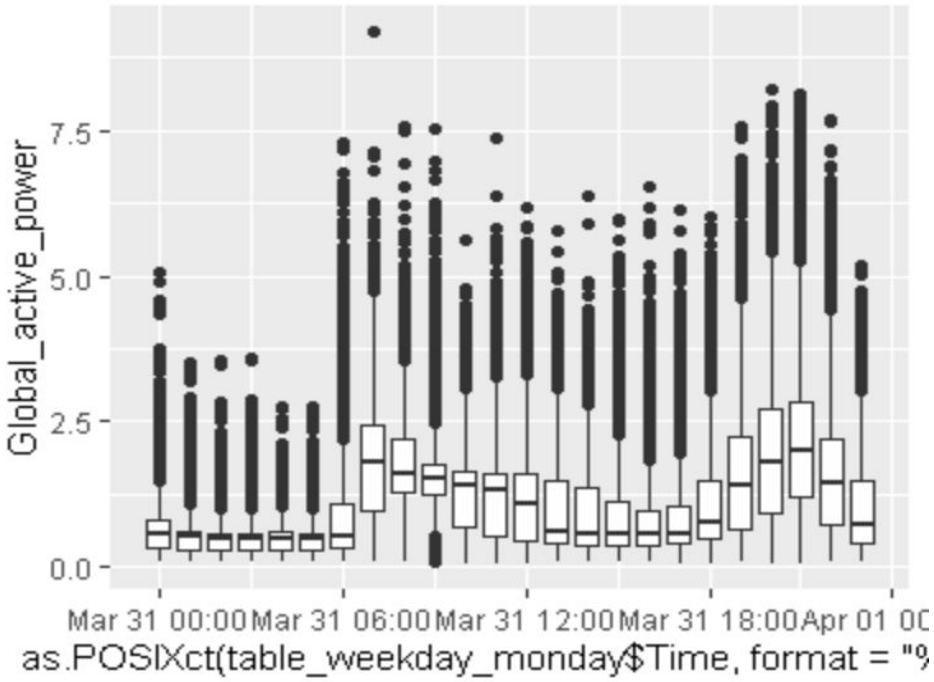
From all three graphs, it can be easily confirmed that Global Active Power & Global Intensity have strong positive correlations and Global Active power & Voltage have strong negative correlation.

By observing the mean displayed in the box plot of the global active power, we can observe two peaks. The time frame where the peaks occur can be easily explained by considering people's usual daily living pattern. The first peak happens at around 8:00 which is when people start using electricity for cooking, hot water, and such to get ready to work. The second peak happens around 20:00 which is when people have already returned home from work (somewhere around 18:00) and start using electricity for the same uses for the first peak. However, the second peak at 20:00 is greater than the first peak (8:00) because at night people will have their lights turned on in their residence.

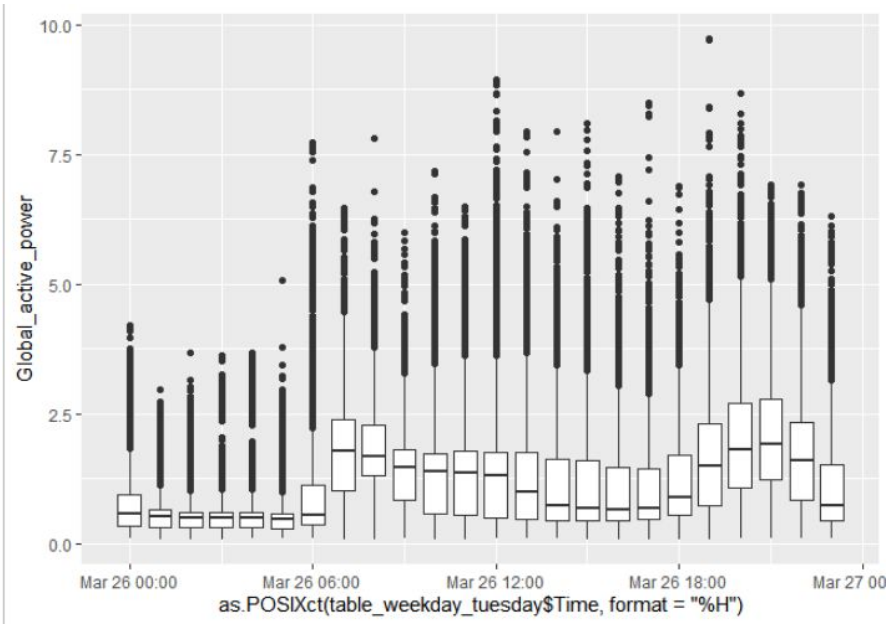
This now leads to the question of whether every day of the week follows this general pattern.

Observing Monday , we can see a clear pattern, having low global active power from 0:00 to 6:00 where people start to wake up. By 7:00 where a lot more people wake up, there is a spike in global active power and it gradually decreases at roughly 10:00 where it stabilizes.

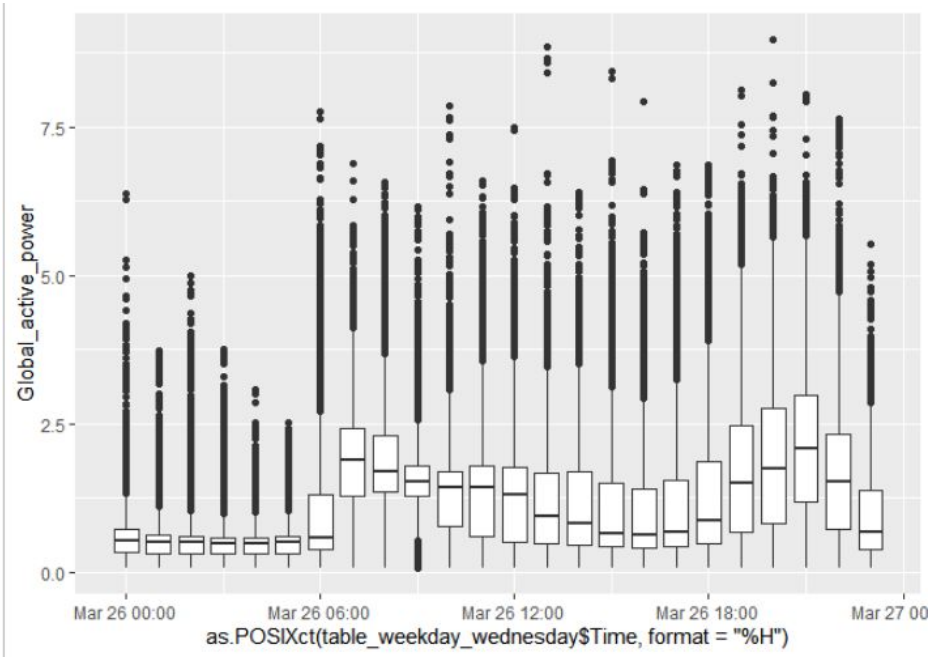
Global Active Power (Monday)



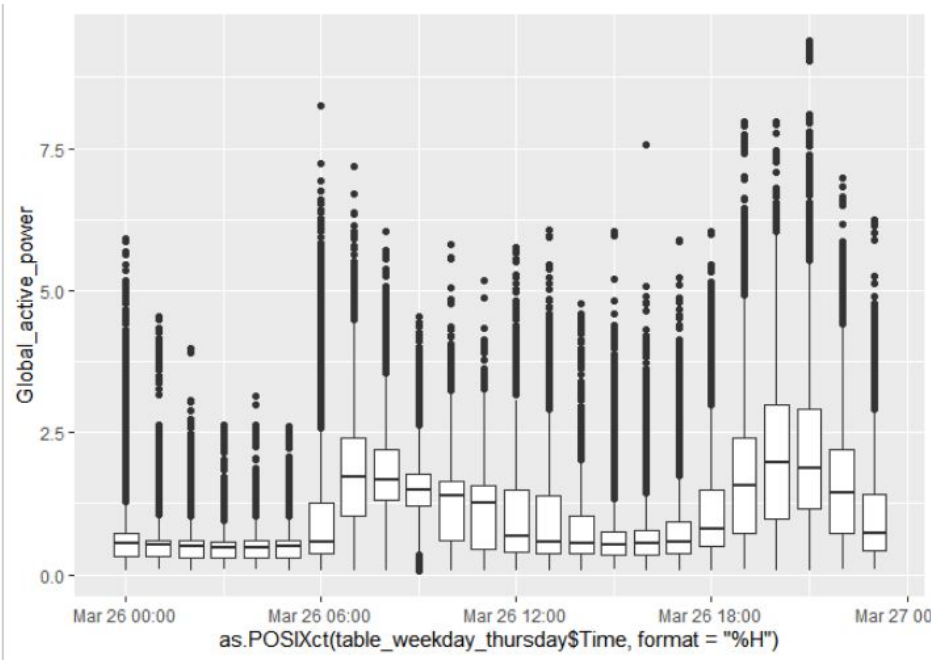
Global Active Power (Tuesday)



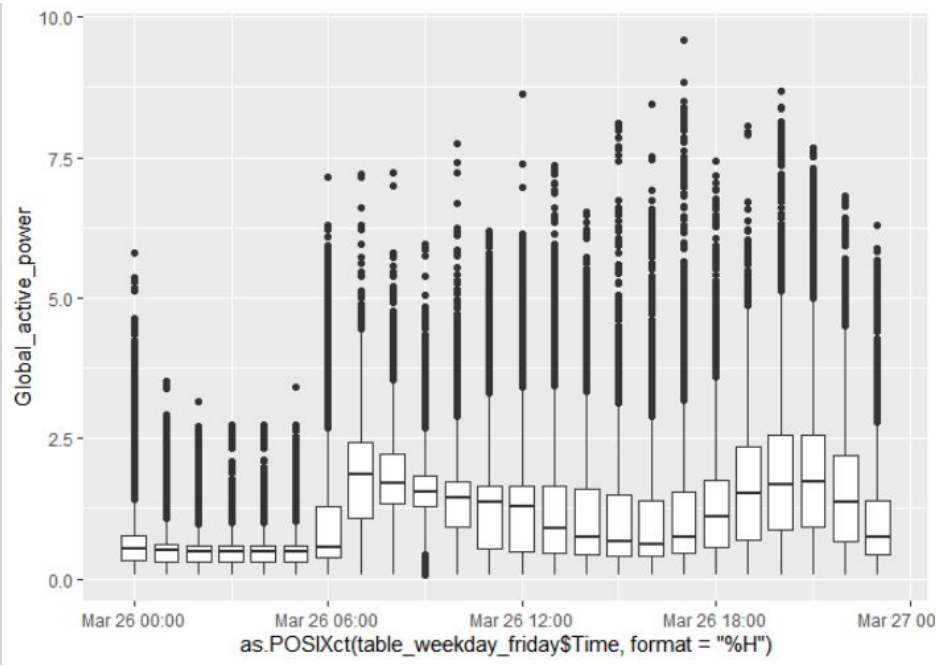
Global Active Power (Wednesday)



Global Active Power (Thursday)

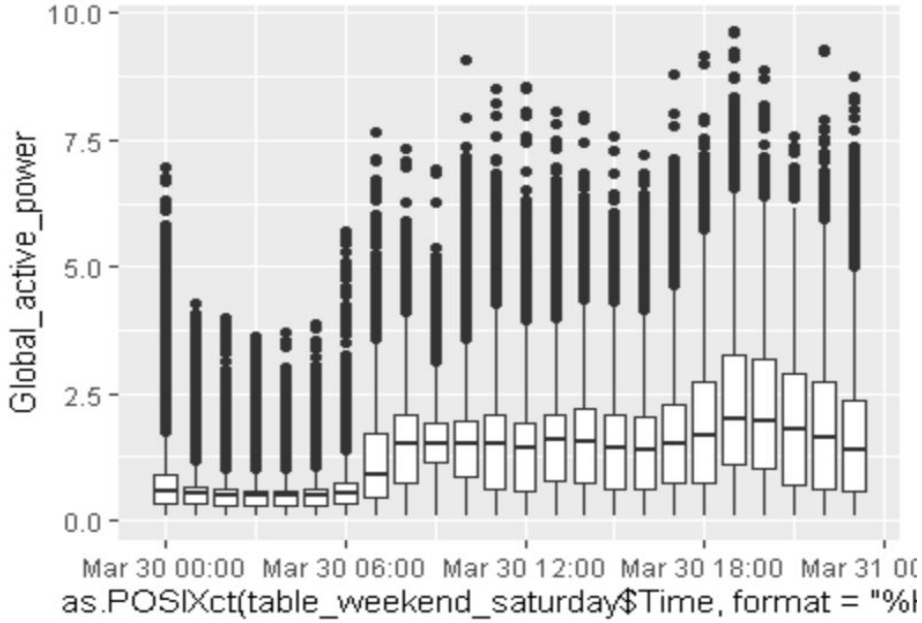


Global Active Power (Friday)

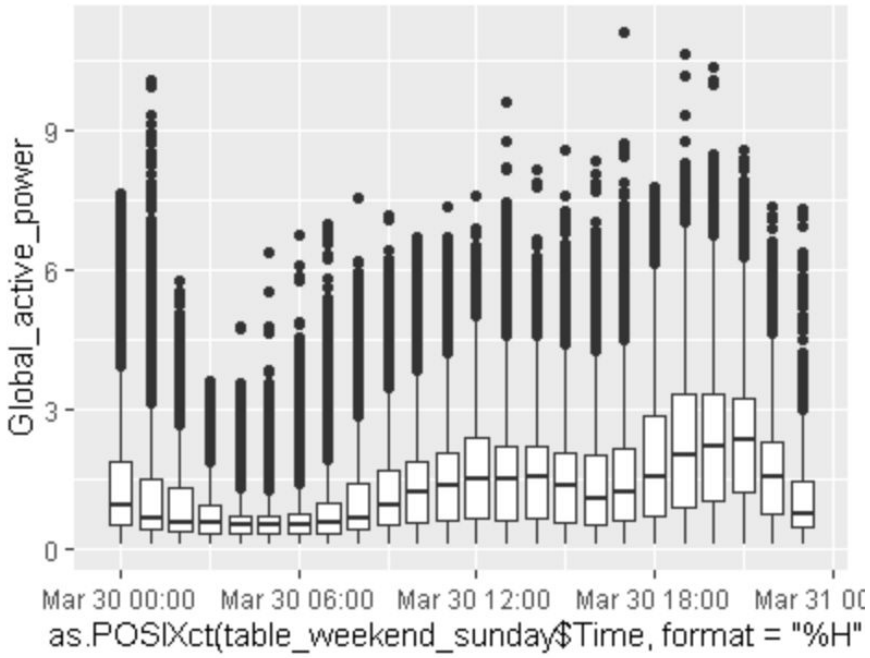


Weekend's have a slightly different pattern. We can see for both weekends, at 0:00 there is a decent usage of global active power and it gradually drops with Sunday having a higher usage as well as a drop. For Saturday, the spike when global active power starts at 7:00 where as Sunday seems to gradually increase till 10:00 with the highest morning spike. Therefore, 7:00 to 10:00 is the chosen time frame.

GLOBAL ACTIVE POWER (SATURDAY)

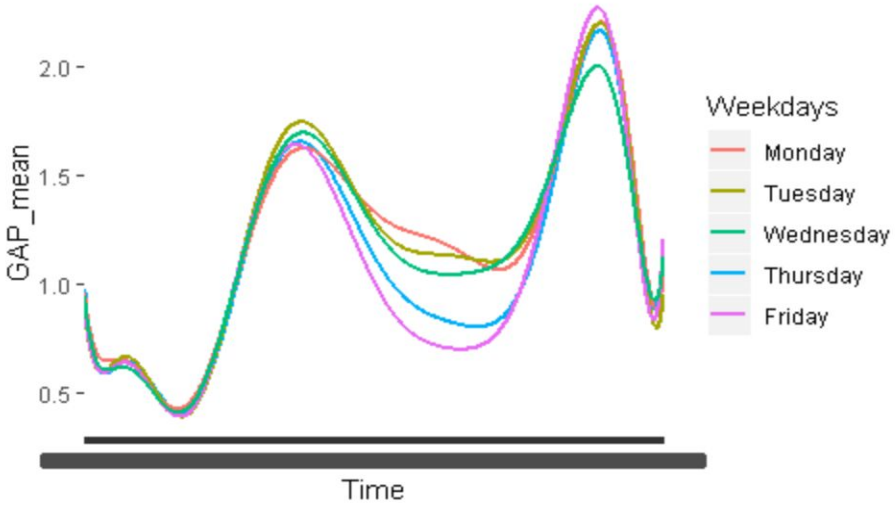


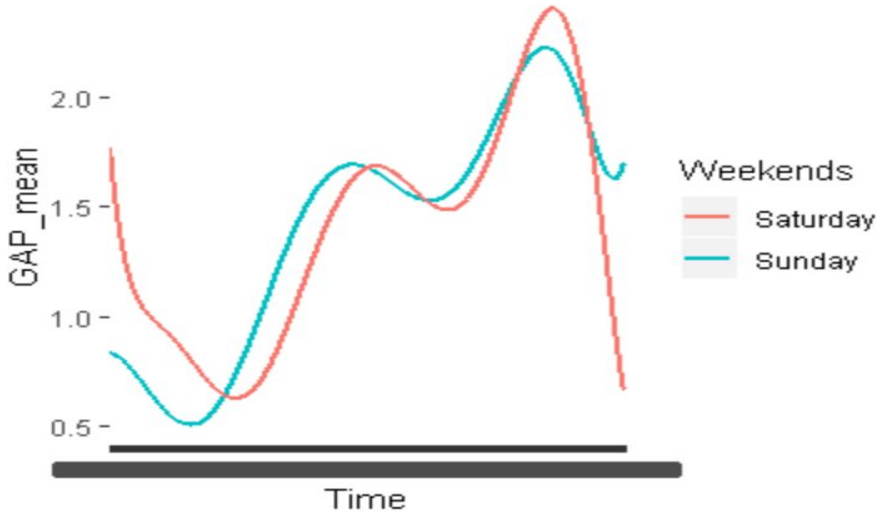
GLOBAL ACTIVE POWER (SUNDAY)



Intuitively, weekdays and weekends have different power consumption data due to the nature of people's working hours. Also, the difference between each weekday and the difference between the two weekend days should be insignificant. The graphs below explain that there is little difference between Monday and Friday and between Saturday and Sunday in terms of global active power.

POLYNOMIAL REGRESSION





Performing polynomial regression on weekdays and weekends, verifies that during 7:00 to 10:00 all weekday's regression lines are very close to each other. We observe that the regression lines for weekends are not as close to one another as the regression line displayed on weekdays but we can recognize the same pattern. As a result, we've selected 7:00 to 10:00 as our time window. Weekdays and weekends

present a relatable pattern throughout the day. Therefore, we've selected Monday as weekday day and Saturday as the weekend day in this data analysis are chosen.

Feature Engineering

For this section, a suitable dependent variable or a combination of Several dependent variables will be chosen for training univariate or multivariate Hidden Markov models. From the Data Exploration section, it was determined that Global Active Power, Global Intensity, and Voltage had highest correlations in the datasets. Also, from plotting them, it was shown that all three response variables display very simple and observable patterns. As a result, there is a high belief that those three response variables are the most suitable to optimize the accuracy. While minimizing the complexity of the models. Principal Component Analysis (PCA) will be used to justify choosing those three response variables.

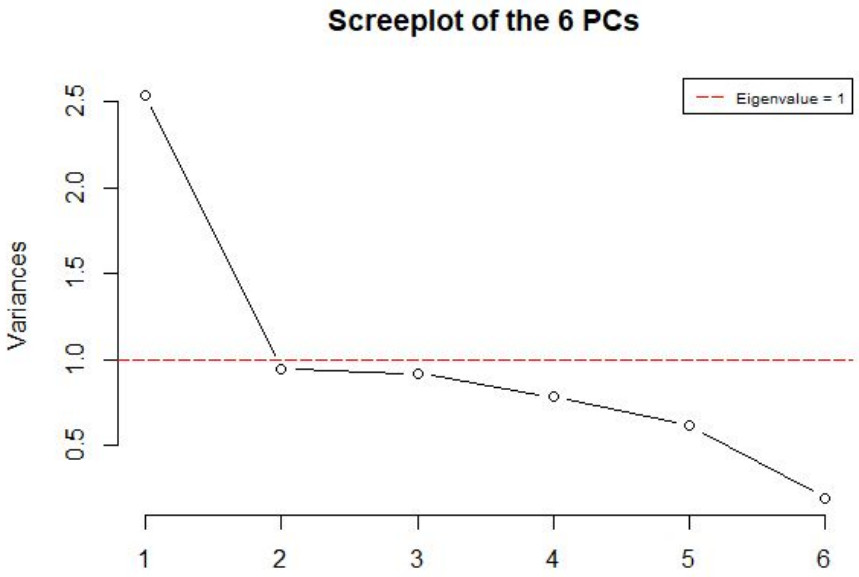
In order to confirm the previous choice of response variable, PCA analysis was made. The table below shows the output of the PCA analysis on the variables Global Active power, Global Intensity, Voltage and Sub Meters 1 - 3. Ideally, the first 3 principle components (Global Active Power, Global Intensity, and Voltage) should retain at least 70-80% of cumulative variance [4]. According to the table, by keeping the

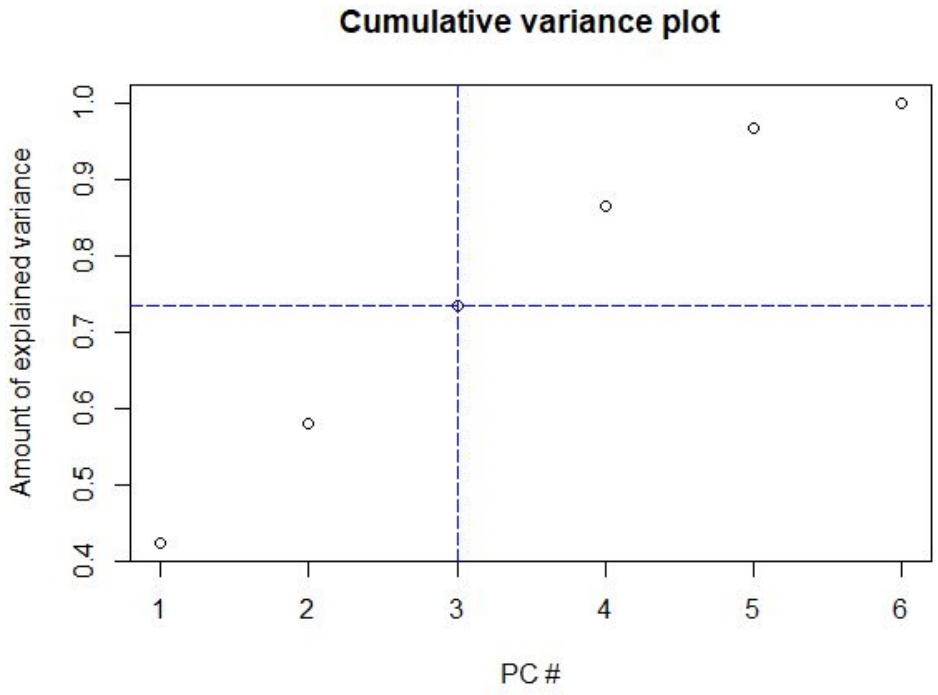
previously mentioned three components and discarding the remaining, the model still can retain 73.39% of cumulative variance.

PRINCIPAL COMPONENTS ANALYSIS DATA

Standard Deviation	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	1.593	0.9609	0.9726	0.8447	0.7859	0.4403
Proportion of Variance	0.423	0.1847	0.1533	0.1308	0.1029	0.0323
Cumulative Proportion	0.423	0.6526	0.7339	0.8648	0.9677	1.0000

Now, the PCA analysis summary was plotted for further confirmation.





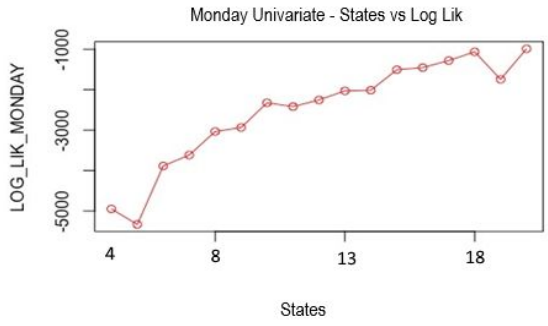
Now, the PCA analysis summary was plotted for further confirmation. Only the first component has Eigenvalue greater than 1. However, as every datas is imperfect, PC2 and PC3 are considered together with PC1 as well since Eigenvalues of PC2 and PC3 are very close to 1 [4]. Therefore, three components should explain approximately 73% of variance. This means that the dimension of the model can be reduced dimension from 6 to 3 and only use about 30 percent of the variance.

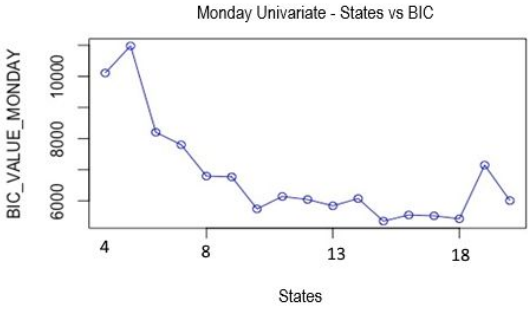
Training and Testing

From Data Exploration, it was determined to choose Monday for weekdays and Saturday for weekends for training. To be able to have a reasonable prediction on future data, statistics is needed to train a model based on the observed data. Hidden Markov Models are the decided type of model to train and test our data. In Markov models, the observer can see the states, and therefore the state transition probabilities are the only parameters, while in the hidden Markov model, the observer cannot see the states, but can see the output observations, dependent on the state [6]. Now the correct number of states must be chosen to create the Hidden Markov Model for univariate (Global Active Power) and multivariate (Global Active Power, Global Intensity, and Voltage), and continue onto the next step which is anomaly detection.

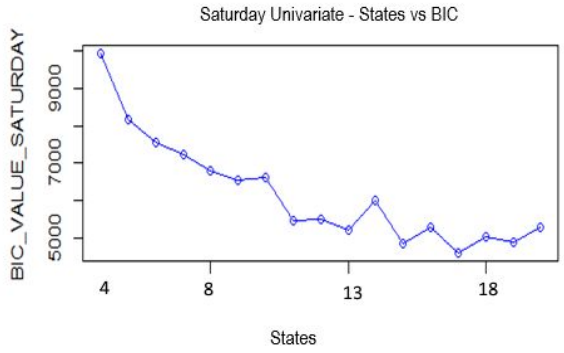
The models from state 4 to 20 will be trained and compare them based on their Bayesian Information Criterion (BIC) and Log-likelihood (Log Lik) to search for the best candidate that can give the most accurate result. The likelihood of an observation sequence is the product of each event's probability in the sequence. Log likelihood is used instead of likelihood because likelihood is a very small number since it is a product of probabilities, numbers between zero and one. The most chosen state is expected to have the lowest BIC and Log-Lik values.

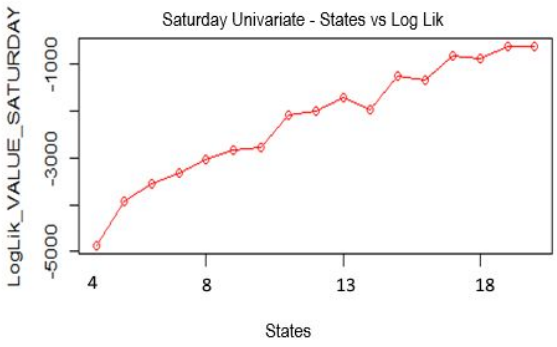
Univariate Training on Monday		
States	BIC	Log Lik
4	10107.91	-4948.301
5	10980.19	-5333.907
6	8202.24	-3885.214
7	7802.882	-3616.629
8	6795.689	-3034.939
9	6769.762	-2934.694
10	5740.377	-2323.533
11	6140.179	-2417.778
12	6041.695	-2253.693
13	5839.793	-2028.711
14	6072.974	-2012.083
15	5346.779	-1506.579
16	5544.204	-1453.699
17	5517.628	-1279.629
18	5422.568	-1062.131
19	7146.029	-1744.706
20	6005.434	-986.0647



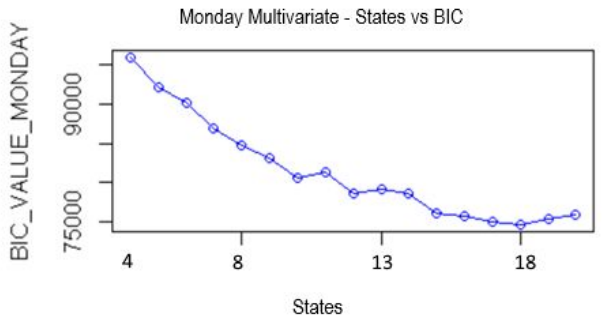


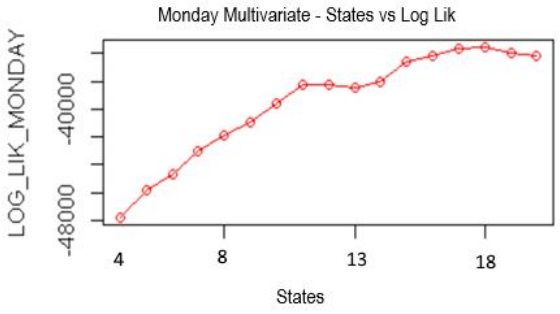
Univariate Training on Saturday		
States	BIC	Log Lik
4	9936.923	-4863.02
5	8163.009	-3925.635
6	7557.848	-3563.458
7	7223.877	-3327.706
8	6786.815	-3031.241
9	6538.339	-2819.899
10	6613.7	-2761.307
11	5460.247	-2079.14
12	5500.859	-1984.835
13	5223.653	-1722.454
14	5997.366	-1976.363
15	4853.481	-1262.304
16	5294.671	-1331.614
17	4604.117	-825.8836
18	5019.093	-863.7489
19	4895.142	-622.9818
20	5264.445	-619.6733



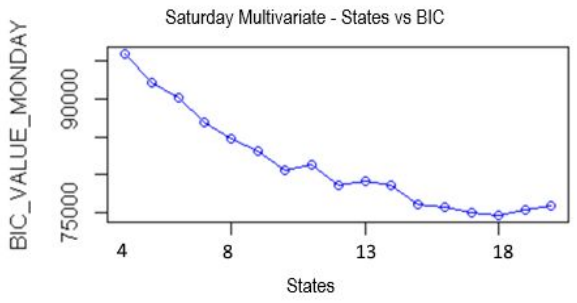


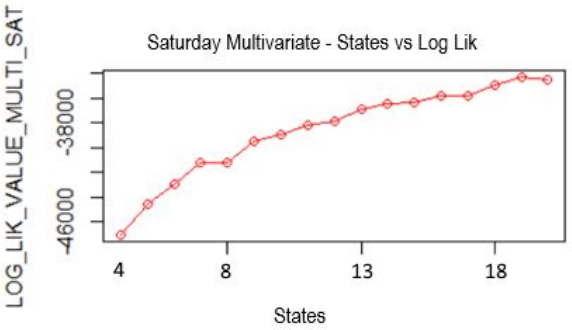
Multivariate Training on Monday		
States	BIC	Log Lik
4	95928	-47785
5	92104	-45804
6	90123	-44735
7	86944	-43058
8	84755	-41867
9	83079	-40924
10	80623	-39581
11	81359	-38289
12	78555	-38289
13	79214	-38477
14	78619	-38028
15	76088	-36601
16	75638	-36206
17	74914	-35665
18	74657	-35536
19	75440	-35928
20	75820	-36118





Multivariate Training on Saturday		
States	BIC	Log Lik
4	94485.47	-47063.94
5	89741.41	-44623.15
6	86515.46	-42932.24
7	83289.5	-41232.16
8	83514.06	-41248.16
9	80310.71	-39541.05
10	79356.17	-38949.17
11	78213.14	-38253.87
12	77701.67	-37865.19
13	76043.46	-36893.97
14	75400.41	-36421.16
15	75589.07	-36355.04
16	74922.45	-35852.1
17	75184.63	-35804.4
18	73908.84	-34978.54
19	72950.12	-34302.06
20	73815.6	-34528.5





After training the four Hidden Markov Models, the graphs reveal that the Monday univariate model best uses 15 states, Monday multivariate model best uses 17 states, Saturday univariate model best uses 18 states, and Saturday multivariate model best uses 19 states because of its BIC values and Log-Lik values respectively.

Chosen States of Monday and Saturday			
	State	BIC	Log-Lik
Univariate Training on Monday	15	5346.779	-1506.579
Univariate Training on Saturday	17	4604.117	-825.8836
Multivariate Training on Monday	18	74657	-35536
Multivariate Training on Saturday	19	72950.12	-34302.06

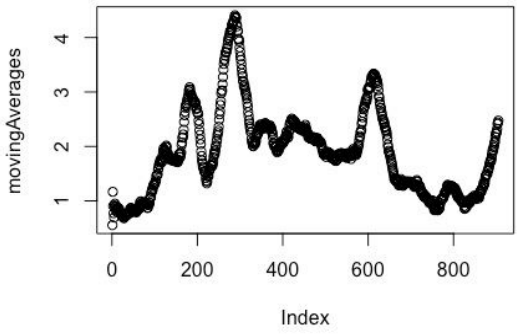
Anomaly Detection

tIn this part, the two methods for anomaly detection are applied - Moving Averages and HMM models chosen in the training process. Since the difference between each week for power consumption is negligible, picking a random week is sufficient as long as that week does close to the mean week. Week 21.02.2010 - 27.02.2010 is the chosen week for applying the moving averages method.

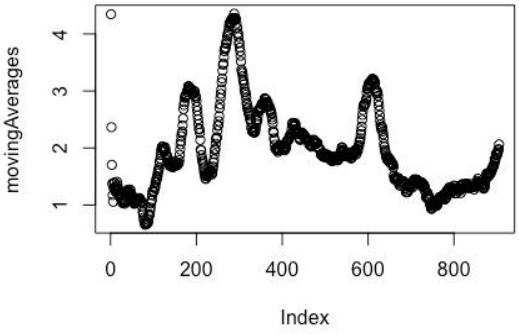
Moving average for weekdays and weekends over the time window from 7am to 10am:

MOVING AVERAGE OF DATASETS

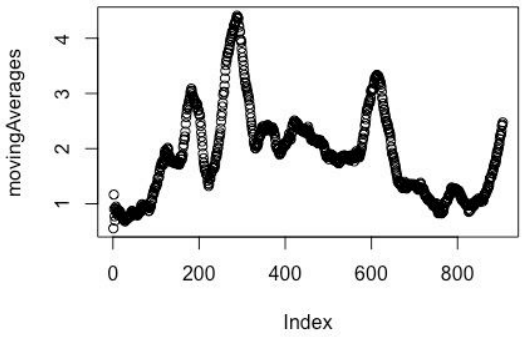
Dataset 1 (weekdays)



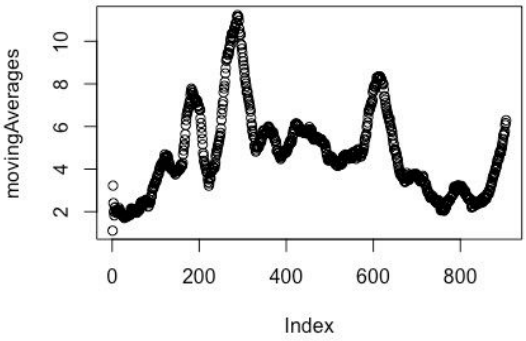
Dataset 2 (weekdays)



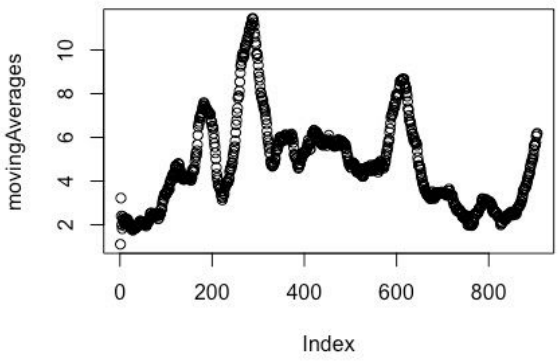
Dataset 3 (weekdays)



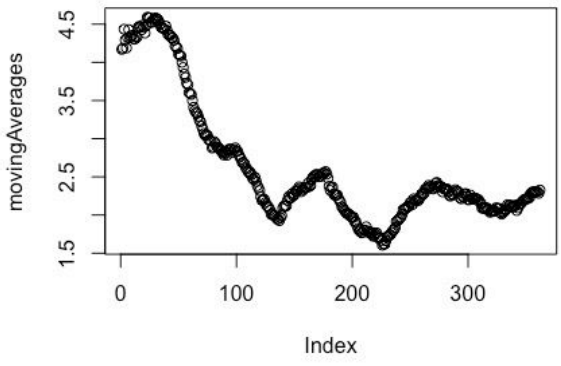
Dataset 4 (weekdays)



Dataset 5 (weekdays)

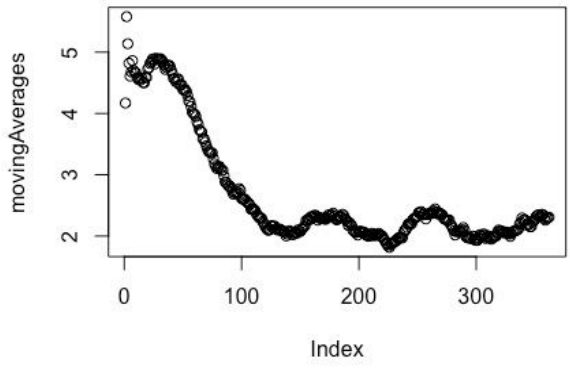


Dataset 1 (weekends)

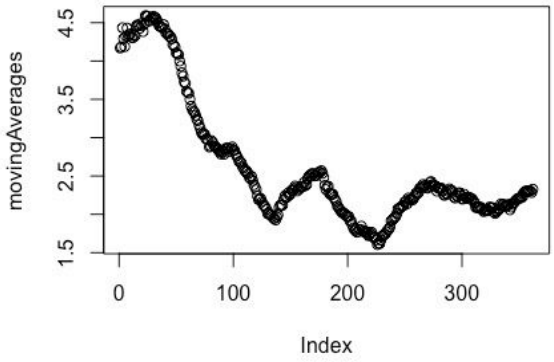


Dataset 2 (weekends)

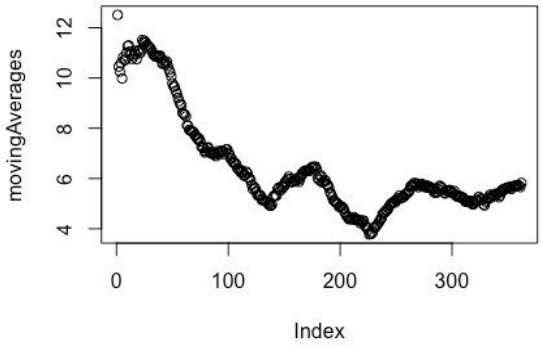
Dataset 3 (weekends)

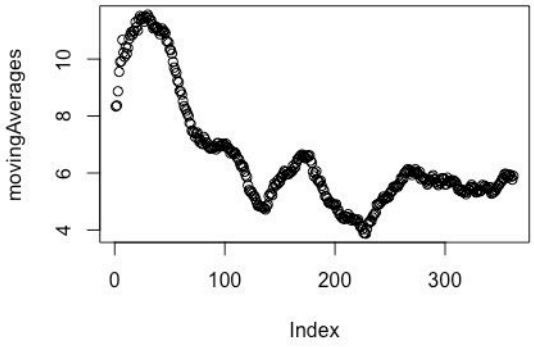


Dataset 4 (weekdays)



Dataset 5 (weekdays)

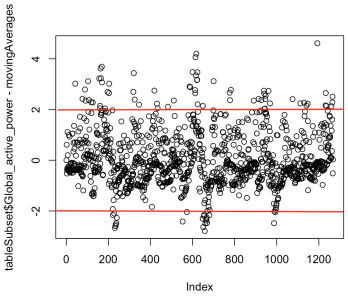




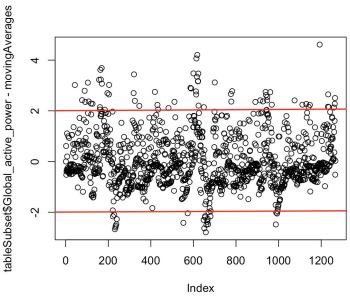
We have computed the moving averages over the time window from 7am to 10am for all five test datasets for the chosen week in order to identify the threshold values computed as a difference of data and moving averages to be used in the anomaly detection process.

THRESHOLD VALUES

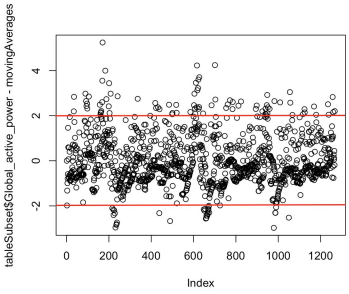
Dataset 1 - Weekdays



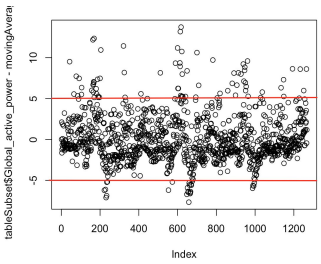
Dataset 2 - Weekdays



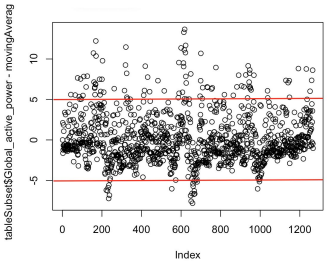
Dataset 3 - Weekdays



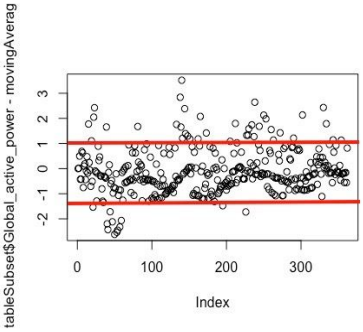
Dataset 4 - Weekdays



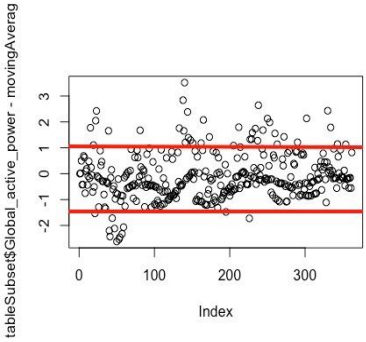
Dataset 5 - Weekdays



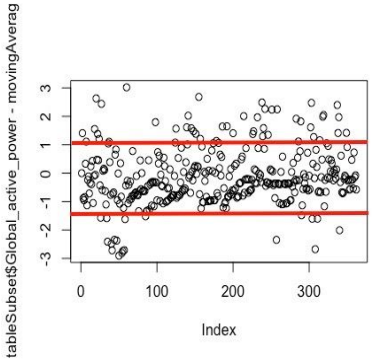
Dataset 1 - Weekend



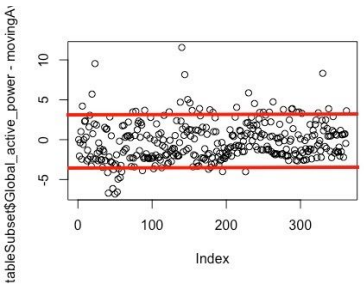
Dataset 2 - Weekend



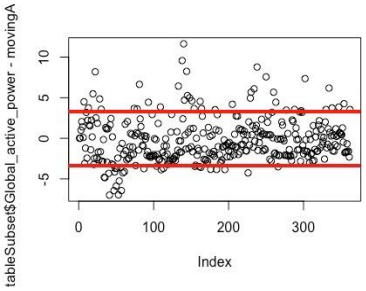
Dataset 3 - Weekend



Dataset 4 - Weekend



Dataset 5 - Weekend



Based on our findings from the moving averages method, the decision has been made to define the Threshold values as:

THRESHOLD VALUES

Dataset	Weekdays	Weekends
Test 1	2 , -2	1, -1.5
Test 2	2 , -2	1, -1.5
Test 3	2 , -2	1, -1.5
Test 4	5 , -5	3, -4
Test 5	5 , -5	3, -4

Now, previously trained Univariate and Multivariate models are used as a basis that shows the expected behaviour, with the best models chosen based on their BIC and likelihood values from the Training Part: Weekdays (Univariate - 15 states, Multivariate - 18 states) and Weekends (Univariate - 17 states, Multivariate - 19 states). The

Log-Likelihood values are computed for all 5 Test dataset data over a period of one year 31.12.2009 - 31.12.2010, also over the chosen time window (7am - 10am).

LOG-LIKELIHOOD FOR ANOMALY DATASETS

Log Likelihood	Univariate - 15 states	Multivariate - 18 states
Weekdays - Test 1	-60379.58	-229479.6
Weekdays - Test 2	-60306.86	-230061.7
Weekdays - Test 3	-60575.37	-232126.1
Weekdays - Test 4	-122187.5	-462810.8
Weekdays - Test 5	-122351.5	-459365.2
	Univariate - 17 states	Multivariate - 19 states
Weekend - Test 1	-17187.25	-46007.82
Weekend - Test 2	-17357.28	-65839.1
Weekend - Test 3	-17306.2	-38384.14
Weekend - Test 4	-35458.04	-133465.2
Weekend - Test 5	-35517.38	-133002.6

Hardships

The raw data contained many corrupt values. Some fields were left blank in the data. Cleaning the data involved removing such values.

With many possible time windows, picking a time zone was a difficulty. The team debated amongst different time windows to capture the desired graph pattern.

Originally, the chosen time window was quite large. This consequently increased the running time to build and train the hidden Markov Models. Training these models took multiple computers to reduce running time. However, after decreasing the time window, the models trained in a reasonable amount of time.

Conclusion

During the project phases, training datas were explored using various statistical methods such as correlations and regression in order to look for relevant response variables and a specific time window that displays recognizable patterns which was 7:00 to 10:00 for Global Active Power, Global Intensity, and Voltage. After this, principal component analysis was used to check for cumulative variance of those three response variables. The result was between 70~80% which shows that the three response variables are correlated enough to use it to create the Hidden Markov Model given the train dataset.

From Data Exploration, it was determined to choose Monday for weekdays and Saturday for weekends for training. Now the correct number of states must be chosen to create the Hidden Markov Model for univariate (Global Active Power) and multivariate (Global Active Power, Global Intensity, and Voltage), and continue onto the next step which is anomaly detection. The models from state 4 to 20 will be trained and compare them based on their Bayesian Information Criterion (BIC) and Log-likelihood (Log Lik) to

search for the best candidate that can give the most accurate result. The most chosen state is expected to have the lowest BIC and Log-Lik values.

Four Hidden Markov models were needed. One univariate and one multivariate for Monday and Saturday each. The Hidden Markov Models for the univariate model included Global Active Power for its variable and the multivariate model included Global Active Power, Global Intensity, and Voltage. After training these models with different numbers of states, the graphs show that Monday univariate model best uses 15 states, Monday multivariate model best uses 17 states, Saturday univariate model best uses 18 states, and Saturday multivariate model best uses 19 states.

For the anomaly detection phase datasets with injected anomalies have been analyzed with the moving average method. For datasets the analyzing week was chosen randomly and the time window was from 7:00 am to 10:00 am. The main analyzed feature was Global Active Power. Moreover, each week was divided into weekdays and weekends and analyzed with a moving average approach. The threshold values were stated for the first three datasets as the gap from -2 to 2 for weekdays and from -1.5 to 1 for weekends. Therefore, any data that lies above or below these values can be considered as an anomaly. For another two datasets, the threshold value gap was chosen from -5 to 5 for weekdays and from -4 to 3 for weekends.

References

[1] J. R. Minkel, “The 2003 Northeast Blackout--Five Years Later,” Scientific American, 13-Aug-2008. [Online]. Available: https://www.scientificamerican.com/article/2003-blackout-five-years-later/.

[2] William Feller. An Introduction to Probability Theory and Its Applications. John Wiley & Sons, 1968.

[3] G. James, D. Witten, T. Hastie, and R. Tibshirani. An Introduction to Statistical Learning: with Applications in R, Springer, 2017.

[4] P. Nistrup, “Principal Component Analysis (PCA) 101, using R,” Medium, 28-Jan-2020. [Online]. Available: https://towardsdatascience.com/principal-component-analysis-pca-101-using-r-361f4c53a9ff.

[5] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286, Feb. 1989.

[6] Lawrence R. Rabiner. A tutorial on hidden Markov models and selected application in speech recognition.Proc. IEEE, 77:257–286, 1989