Electrical Grid Cybersecurity:
An Anomaly Detection Approach to Examine Abnormal Behaviour and Potential Security Intrusions

CMPT 318 – Cybersecurity
Spring 2020
Prof. Uwe Glässer

Project Group #2

ABSTRACT

Nowadays, many time series are generated by various types of critical infrastructure such as electrical power grids, thermal plants, public water utilities, etc. This report attempts to address the procedures and findings in exploration of available data from electricity consumption of households in the U.S. and the implementation of anomaly detection techniques to determine point and contextual anomalies in five sets of unseen data. Point anomalies are detected by applying the so-called moving average method. Contextual anomalies are determined by trained hidden Markov models (HMMs) and comparison of log-likelihoods to determine anomalous behaviour. For feature selection, a subset of dependent electrical variables is selected based on correlation coefficients and performing a Principal Component Analysis (PCA) on the provided training data set. The analysis and methodologies for the feature engineering, HMM training and testing as well as anomaly detection is addressed in detail within the report.

TABLE OF CONTENTS

TABLE OF FIGURES

1.0. INTRODUCTION

"In light of increasing cyber threats, especially advanced persistent threats (APT), and existing vulnerabilities that expose critical infrastructure to a variety of adversarial scenarios, the project explores behaviour based intrusion detection methods used for cyber situational analysis of automated control procedures (Course Project Document)."

The problem being addressed as part of the CMPT 318 final course project is to develop an anomaly detection approach to detect intrusion and abnormal behaviour of electricity consumption given a number of data sets and the analysis techniques and methodologies that was taught through the course. To approach the problem, the provided training data set is studied to learn and better understand the underlying trends of electrical behaviour in certain time windows, and the significance of the different electrical characteristic features by statistical analysis. Using the training data, an attempt to find anomalies in the given test data is performed using two different methods: the first approach aims to find point anomalies, while the second approach aims to detect contextual anomalies by training a number of hidden Markov models (HMM)—which is a novel behaviour-based intrusion detection method. Throughout the project, team members used built-in R libraries to process the raw data, create effective models and test them for optimal accuracy and effective analysis of anomaly detection.

2.0. BACKGROUND INFORMATION

Due to the exponential progress towards the Internet of Things[1] (IOT) and intensification of automation, cyberattacks are increasingly routine and sophisticated. Automation is vital for the continuous operation of critical infrastructure[2] and the services it provides because it enhances cost, efficiency, quality

[1] A system of interrelated computing devices, mechanical and digital machines: these objects are provided with unique identifiers (UIDs) and the ability to transfer data over a network without requiring human-to-human or human-to-computer interaction.
[2] Refers to processes, systems, facilities, technologies, networks, assets and services essential to the health, safety, security or economic well-being of people and the effective functioning of government.

of service delivery and safe operation of critical assets. However, increasing reliance on automation, and having a large system of interrelated computing devices in effect increases the attack surface for advanced persistent threats[3] and amplifies the risk of cascading effects.

The threat landscape has evolved to the point that risks that were once considered unlikely began occurring with regularity. This ongoing trend can be attributed to higher maturity of attack tools and methods, increased exposure, increased motivation of attackers, and better detection tools enabling more visibility.

In recognition of the limitations of signature-based analysis of network traffic, defenders have shifted their focus to more behavioural-based anomaly detection. This is because signature-based detection schemes contain a time lag between the initial attack and availability of either a signature or a security patch. Furthermore, behaviour-based anomaly detection could overcome the limitation of being able to alert only on specific signature mismatches, and they have the potential to discover evidence of zero-day exploitation[4] through identifying unusual behaviours.

The electrical grid is composed of a highly diverse set of assets, systems, and functions, and is primarily owned and operated by the private sector in the United States or by Provincial, Territorial, investor-owned, and municipal utilities in Canada. In part, because of its complexity and physical size, and the increasing use of networked Industrial Control Systems, the electrical grid is vulnerable to disruptions from a variety of hazards and threats. Enhancing response and recovery efforts depends on collaboration with all stakeholders. The challenge is addressing the continued evolution of physical threats, technological risks, cyber incidents, and natural hazards, including climate change (National Electric Grid Security and Resilience Action Plan, 2016).

Protecting today's electricity grid and enhancing preparedness is a key prerequisite for the effective functioning of each country's economy and the well-being of their citizens. This shows the importance of

---

[3] A set of stealthy and continuous computer hacking process often orchestrated by a person or group targeting a specific entity (either a private organization or a state for both business and political motives).
[4] A cyber-attack that occurs on the same day a weakness is discovered in software. At that point, the attack is exploited before a fix becomes available from the software's owner.

cybersecurity in electrical power grids. Electrical grid security refers to the activities that utilities, regulators, and other stakeholders play in securing the national electricity grid. The American electrical grid is going through one of its largest changes in history, which is the move to smart grid technology. The smart grid allows energy customers and energy providers to manage and generate electricity more efficiently. Similar to other new technologies within the cyberspace, the smart grid also introduces new concerns about security.

The North American electrical power grid is a highly connected system. The ongoing modernization of the grid is generally referred to as the "smart grid". Reliability and efficiency are two key drivers of the development of the smart grid. A key attribute is the ability for the electrical system to incorporate renewable energy sources such as wind power and geothermal power. One of the key issues for the power grid security is that these ongoing improvements and modernizations have created more risk to the system. An example of is such risks being imposed by the integration of digital communications and computer systems with the existing physical infrastructure of the power grid.

As of 2006, over 200,000 miles of transmission lines that are 230 kV or higher existed in the United States. The main problem is that it is impossible to secure the whole system from intrusions and attacks. However, the scenario of such attacks occurring would be minimal because it would only disrupt a small portion of the overall grid. For instance, an attack that destroys a regional transmission tower would only have a temporary impact. The modern-day power grid system is capable of restoring equipment that is damaged by natural disasters such as tornadoes, hurricanes, ice storms, and earthquakes in a short period of time. This is due to the resilience of the national grid to such events. "It would be difficult for even a well-organized large group of terrorists to cause the physical damage of a small- to moderate-scale tornado (Electric utility responses to grid security issues, 2006)".

In 2016, members of the Russian hacker organization "Grizzly Steppe" infiltrated the computer system of a Vermont utility company, Burlington Electric, exposing the vulnerability of the nation's electric grid to attacks. The hackers did not disrupt the state's electric grid, however. Burlington Electric discovered malware code in a computer system that was not connected to the grid (Russian operation hacked a Vermont utility, showing risk to U.S. electrical grid security, officials say, 2016).

Based on the above background information and the introduction to risks involved within power grid cybersecurity, it is important to find and address solutions to prevent and mitigate possible impact of electricity grid disruption due to cyber activities.

3.0. METHODOLOGIES AND RESULTS

3.1. DATA EXPLORATION

The primary purpose in data exploration is to identify a particular time window exhibiting clearly recognizable electricity consumption 'patterns of interest' (in the main dependent variable) and to select significant characteristic features in order to create a suitable probabilistic model that could represent a sample 'normal' behaviour for training HMMs.

A data set of approximately three years is provided with the information on electricity consumption from December 16th, 2006 to December 1st, 2009. The data set is order by date and time—providing minute-by-minute data—with characteristic features such as, Global Active Power, Global Reactive Power, Global Intensity and voltage to represent household electricity consumption.

For finding the main variables among different features within the data set, correlation coefficient of all dependent variables is graphed and analysed. **Figure 1** shows the correlation coefficient plot—which is a visualization of the strength, direction, and form of the relationship between each pair of quantitative variables. In addition to **Figure 1**, the correlation coefficient values are recorded in **Table 1**.

	Global_active_power	Global_reactive_power	Voltage	Global_intensity	Sub_metering_1	Sub_metering_2	Sub_metering_3
Global_active_power	1.0000000	0.14035053	-0.2970133	0.7240937	0.28264025	0.29689389	0.44478931
Global_reactive_power	0.1403505	1.00000000	-0.1145850	0.2629112	0.13431478	0.14162172	0.08028465
Voltage	-0.2970133	-0.11458500	1.0000000	-0.4177546	-0.19881568	-0.17445357	-0.28521663
Global_intensity	0.7240937	0.26291116	-0.4177546	1.0000000	0.48796634	0.44917074	0.61634644
Sub_metering_1	0.2826402	0.13431478	-0.1988157	0.4879663	1.00000000	0.05895919	0.10751741
Sub_metering_2	0.2968939	0.14162172	-0.1744536	0.4491707	0.05895919	1.00000000	0.08890123
Sub_metering_3	0.4447893	0.08028465	-0.2852166	0.6163464	0.10751741	0.08890123	1.00000000

Table 1 - correlation coefficient values among different features of the dataset

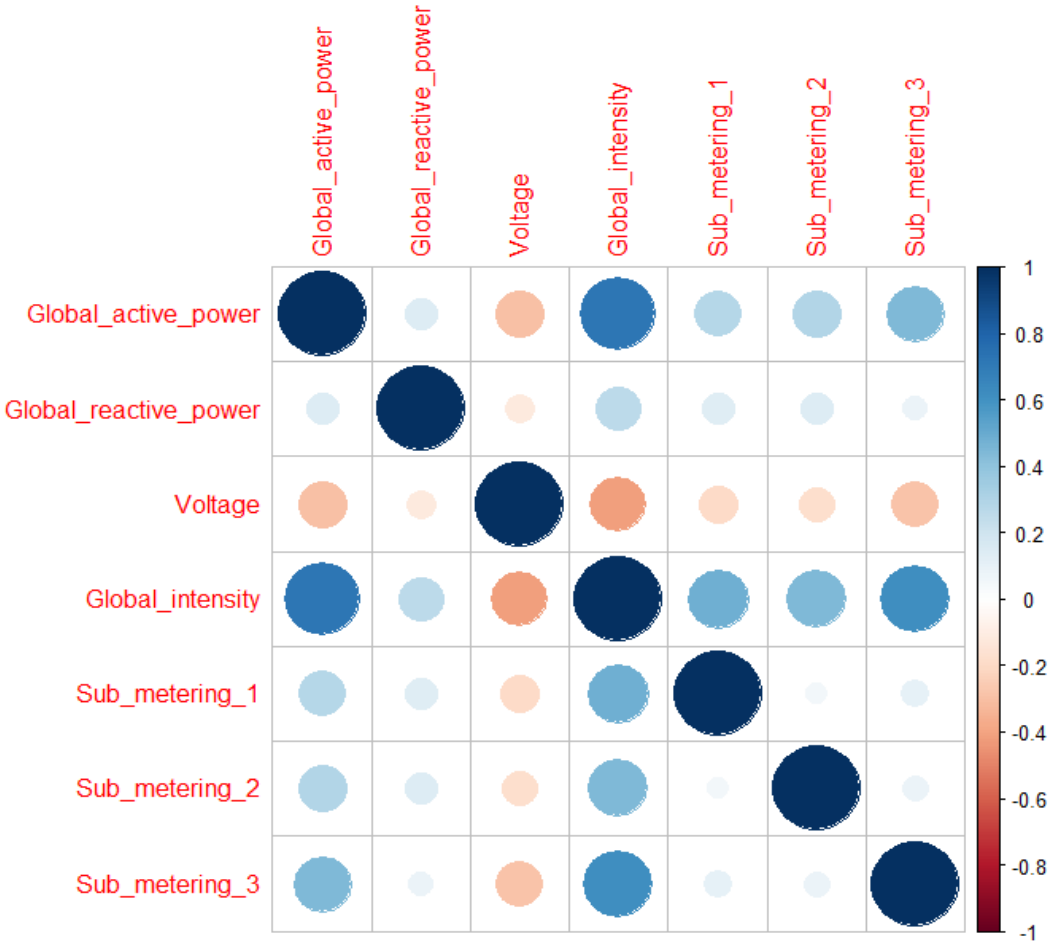


Figure 1 - Plot of correlation coefficients among different features of the data set

For reference, the details about the parameters in the data set as well as available main features are as follows:

- **Date:** Date in format dd/mm/yyyy
- **Time:** Time in format hh:mm:ss
- **Global Active Power:** Household global minute-averaged active power (in kilowatts)
- **Global Reactive Power:** Household global minute-averaged reactive power (in kilowatts)
- **Voltage:** Minute-averaged voltage (in volts)

Based on the results above, it is observed that the correlation between Global Active Power and Global Reactive Power is weak ($r = 0.140$) while Global Active Power and Global Intensity have a strong linear relationship ($r = 0.724$). It is also observed that Voltage is negatively correlated to Global Active Power and Reactive Power with weak values ($r = -0.297$ and $r = 0.262$ respectively) while it has a strong negative correlation with Global Intensity ($r = 0.418$).

As global active power represents the real electrical resistance power consumption in circuits and electrical appliances, data exploration of the provided 'training' data set is performed by hourly aggregation of global active power values. **Figure 2** on this page shows such graph for 24 hours on all weekdays and weekend days. Based on this graph, data aggregation and analysis are

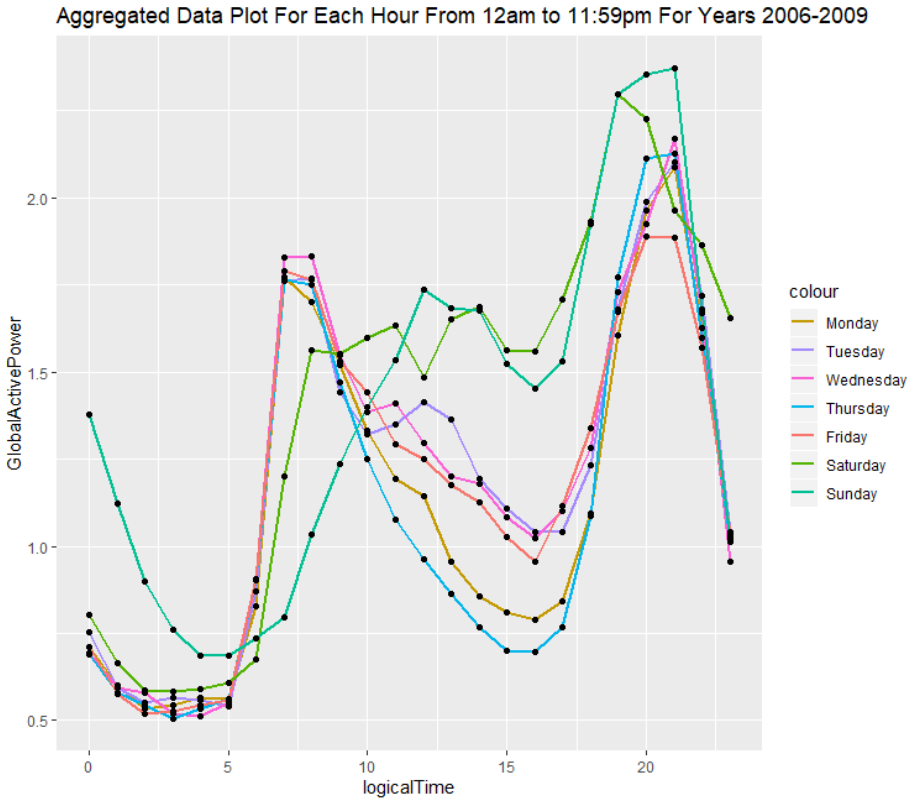


Figure 2 – Hourly aggregated global active power data, plotted for 24 hours for calendar years 2006-2009

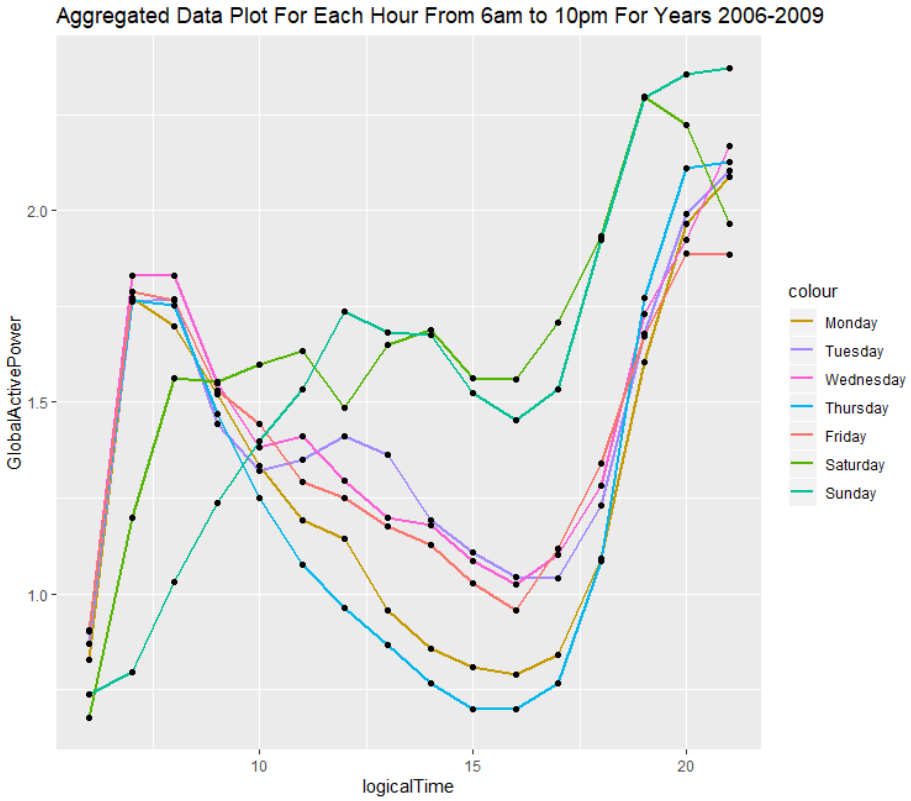


Figure 3 – Hourly aggregated global active power data, plotted from 6am-10pm for calendar years 2006-2009

performed and visualized in **Figure 3** for a shorter time interval (from 6am-10pm) where multiple patterns of interest are observed. To help with the selection of the time window, the graphs are broken down for individual years in **Figure 4**, **Figure** and **Figure** . Note that the timeseries data for years 2006-2007 are combined as 2006 includes data for just over 2 weeks and no useful observable trend can be extracted.

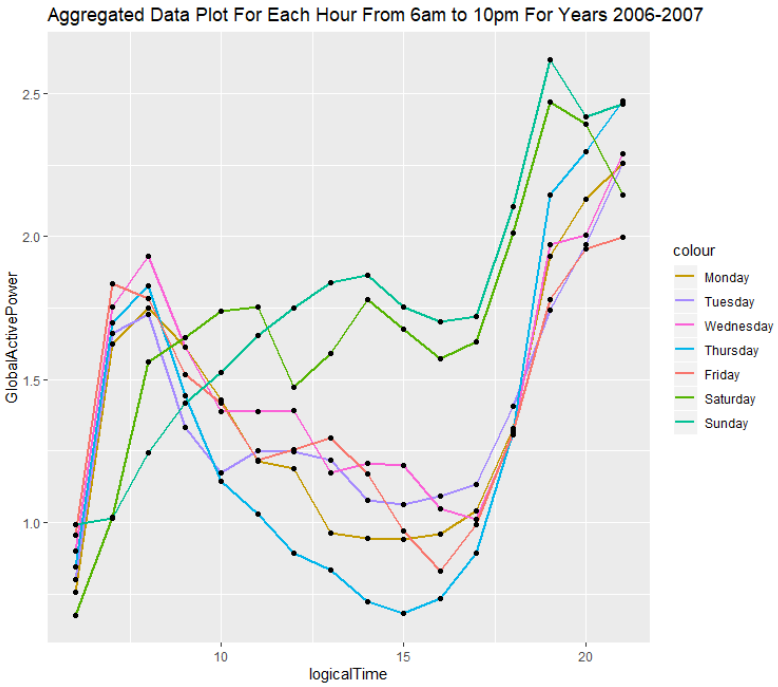


Figure 4 - Hourly aggregated global active power data, plotted from 6am-10pm for calendar years 2006-2007

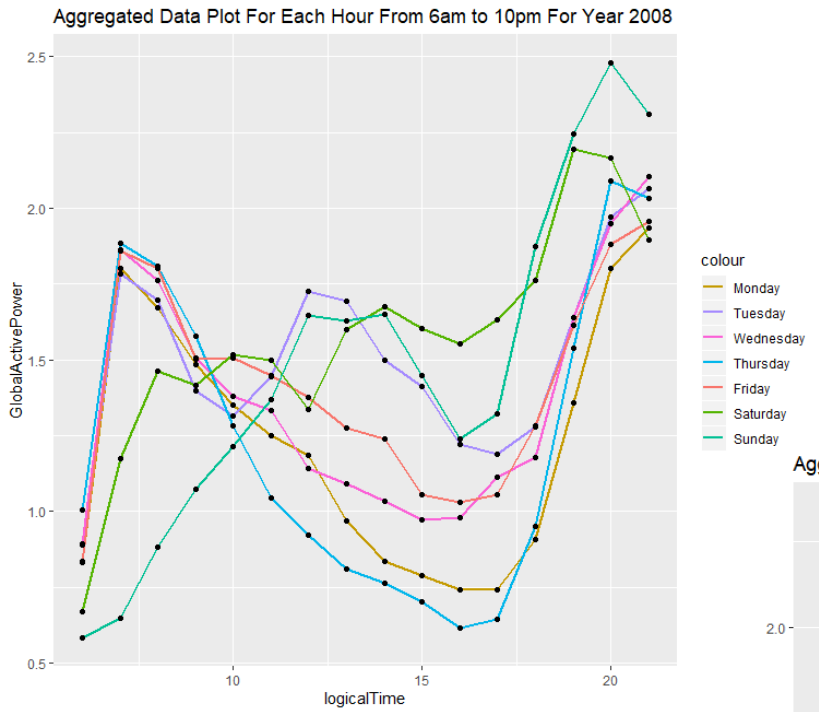
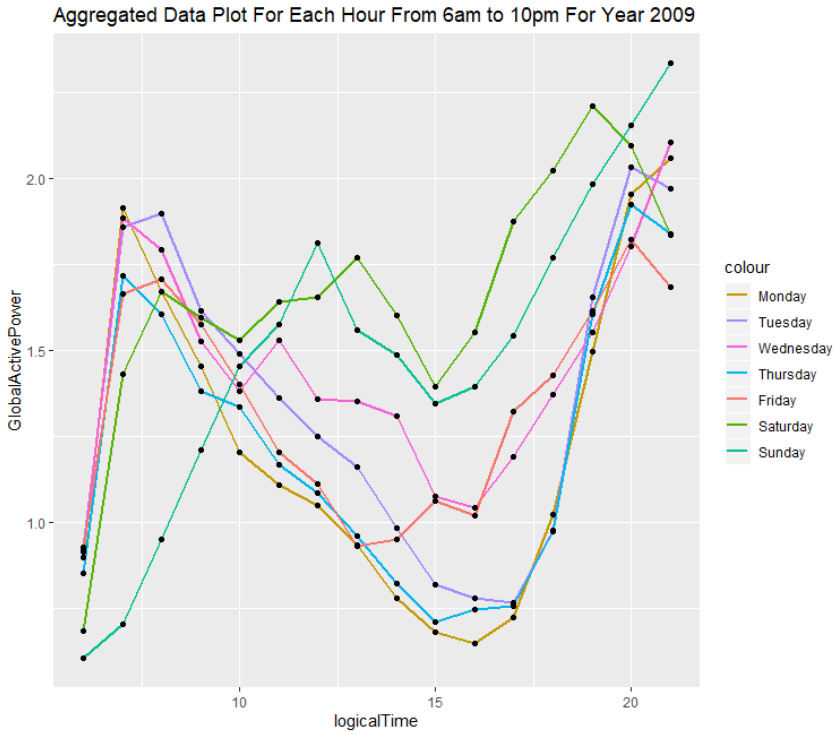


Figure 5 - Hourly aggregated global active power data, plotted from 6am-10pm for calendar years 2008

Figure 6 - Hourly aggregated global active power data, plotted from 6am-10pm for calendar years 2006-2009



Based on analysis of observations on different time intervals within the 4-year aggregated data plot and the aggregated data plots for each year, a selection on 2 different time intervals—9am to 1pm for day time as well as 6pm to 10pm for night time—is made with 3 different weekdays—Mondays, Wednesdays and Fridays, to be narrowed to one day—and 1 weekend day—Sundays—as the observed behaviour of the global active power data follow a general increasing/decreasing trend within the given time frames on all 3 year-by-year graphs and the combined 4-year graph.

For clarity and conciseness, the aggregations and graphs in **Figures 7** and **8** represent the 4-year aggregated data.

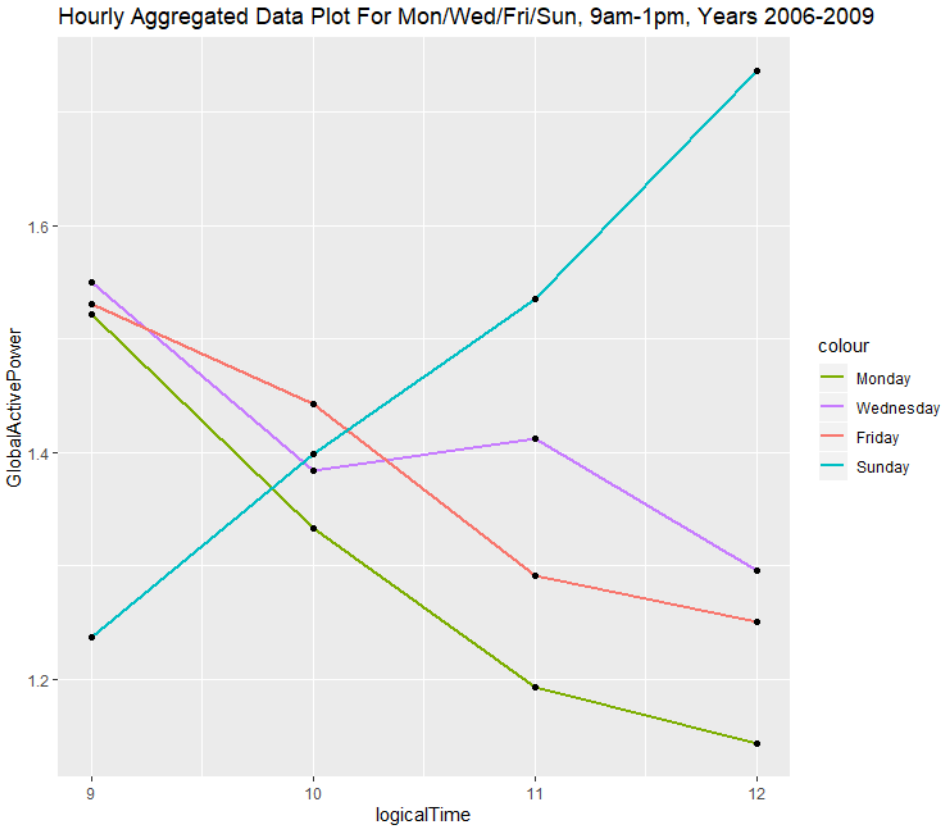


Figure 7 - Hourly aggregated global active power data, plotted from 9am-1pm for calendar years 2006-2009. Note that global active power follows a (generally) decreasing trend on weekdays (analyzed hourly) and an increasing trend on weekends (again, analyzed hourly)

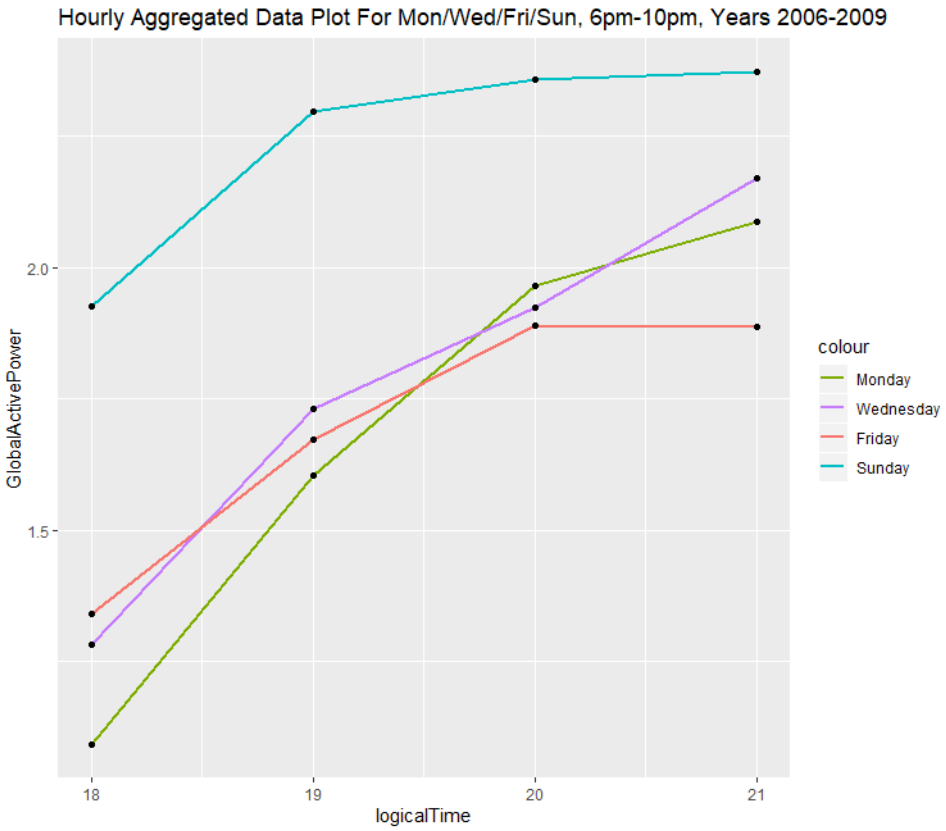


Figure 8 - Hourly aggregated global active power data, plotted from 6pm-10pm for calendar years 2006-2009. Note that global active power follows a (generally) increasing trend on both weeknights and weekend nights (analyzed hourly)

Finally, as the variation in activities during night-time differs for every household, the team selected 9am-1pm during daytime as the single observation time window during weekdays and weekend days. The selected time window generally corresponds to part of work/school hours during weekdays as well as waking-up and leisure routines on weekends. For the specific weekday, the team selected Mondays for further data analysis, HMM modelling and anomaly detection. Graphs below represent the trend in minute-by-minute aggregation of the global active power data and trends for the selected time window on weekdays (**Figure 9**) and weekends (**Figure 10**).

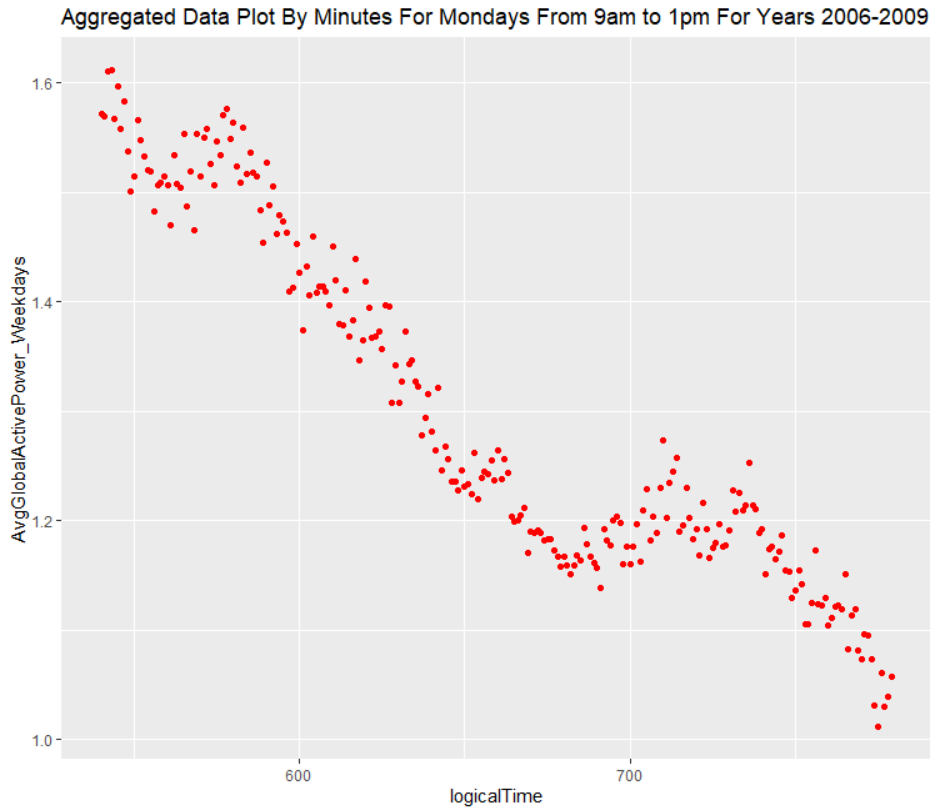


Figure 9 – Minute-by-minute aggregation of global active power data on Mondays (Weekdays), plotted from 9am-1pm for calendar years 2006-2009.

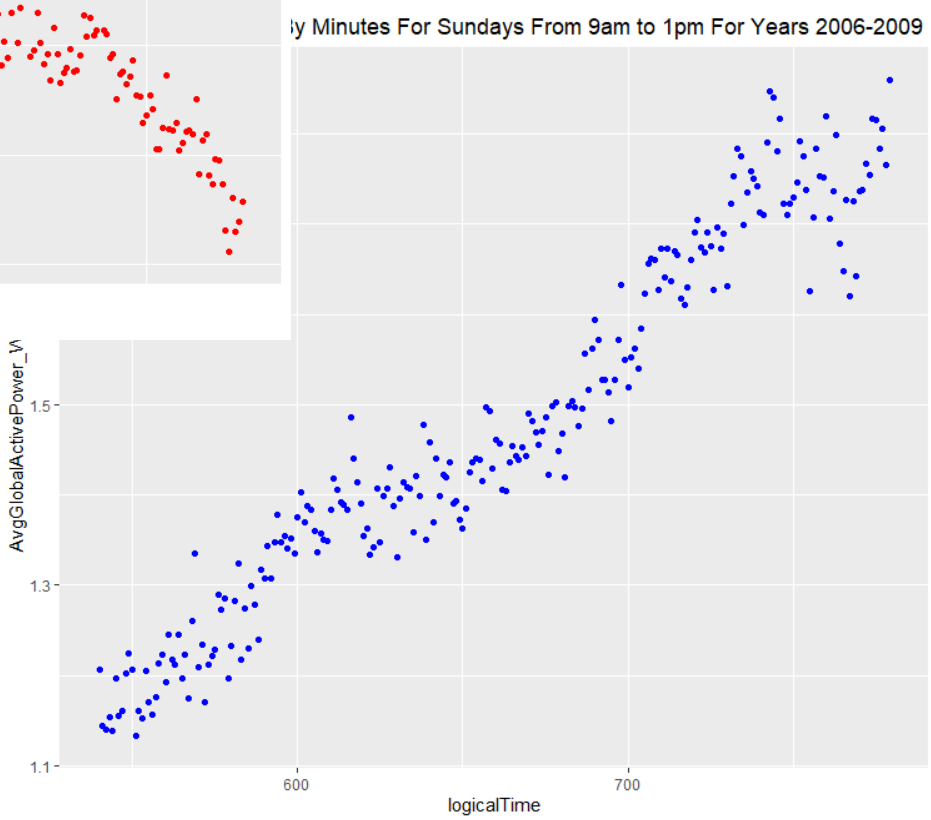


Figure 10 - Minute-by-minute aggregation of global active power data on Sundays (Weekends), plotted from 9am-1pm for calendar years 2006-2009.

3.2. FEATURE ENGINEERING

For training of the univariate and multivariate hidden Markov models, it is required to determine a set of dependent variables from the normal electricity consumption data. In addition, the correlation information needs to be analysed in order to decide which variables are most suitable for HMM modelling and implementation of the Principle Component Analysis[5] (PCA).

3.2.1. COMPUTATION OF PRINCIPAL COMPONENTS

As provided, the 'training' data set contains timestamps (date and time) along with 7 other features. PCA works best with numerical variables, so it is necessary to exclude the categorial features, which include Submetering 1, 2 and 3. For PCA analysis, the following 4 features are considered: Global Active Power, Global Reactive Power, Voltage, Global Intensity. The tools and methods required for PCA analysis are included in 'ggbiplot' package in R.

PCA provides 4 principal components: PC1, PC2, PC3 and PC4. These are applied to the available data window the time window of 9am-1pm for all years in the data set to observe correlation and dependency between the variables. **Table 2** summarizes the percentage of the total variation in the data set.

```
Importance of components:
                          PC1    PC2    PC3     PC4
Standard deviation     1.3503 0.9764 0.9406 0.58188
Proportion of Variance 0.4558 0.2384 0.2212 0.08465
Cumulative Proportion  0.4558 0.6942 0.9153 1.00000
```

Table 2 – summary of the percentage of variation in PCs within the data set

[5] Principal Component Analysis (PCA) is a useful technique for exploratory data analysis, allowing the analysts to better visualize the variations within a dataset with many variables present. It is particularly helpful in the case of "wide" datasets, where we have many variables for each sample. PCA allows us to see the overall "shape" of the data, identifying which samples are similar in characteristics and which are very different.

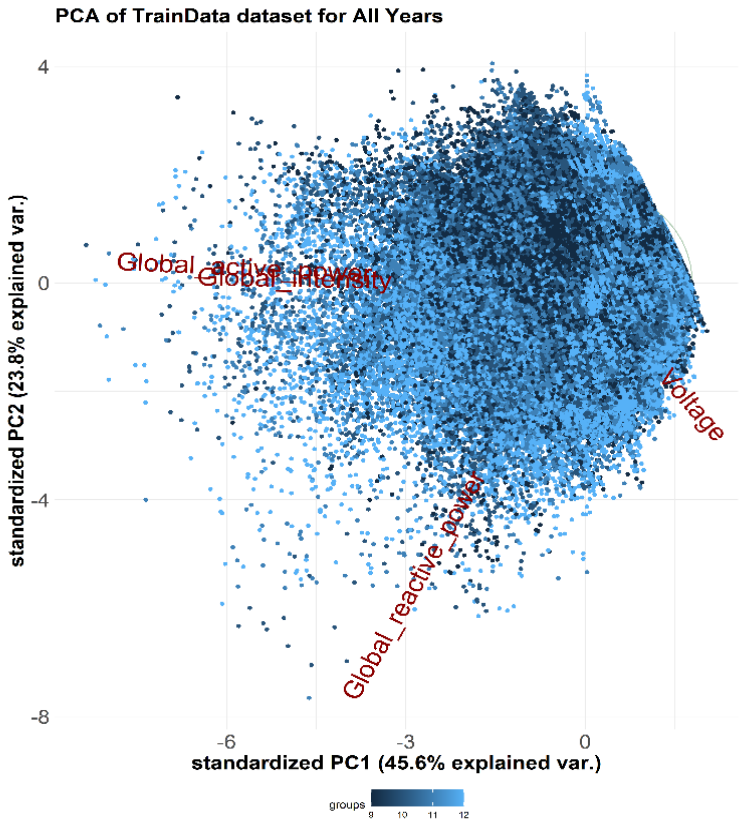


Figure 11 – Distribution of PCs for time window 9am-1pm (aggregated for all years)

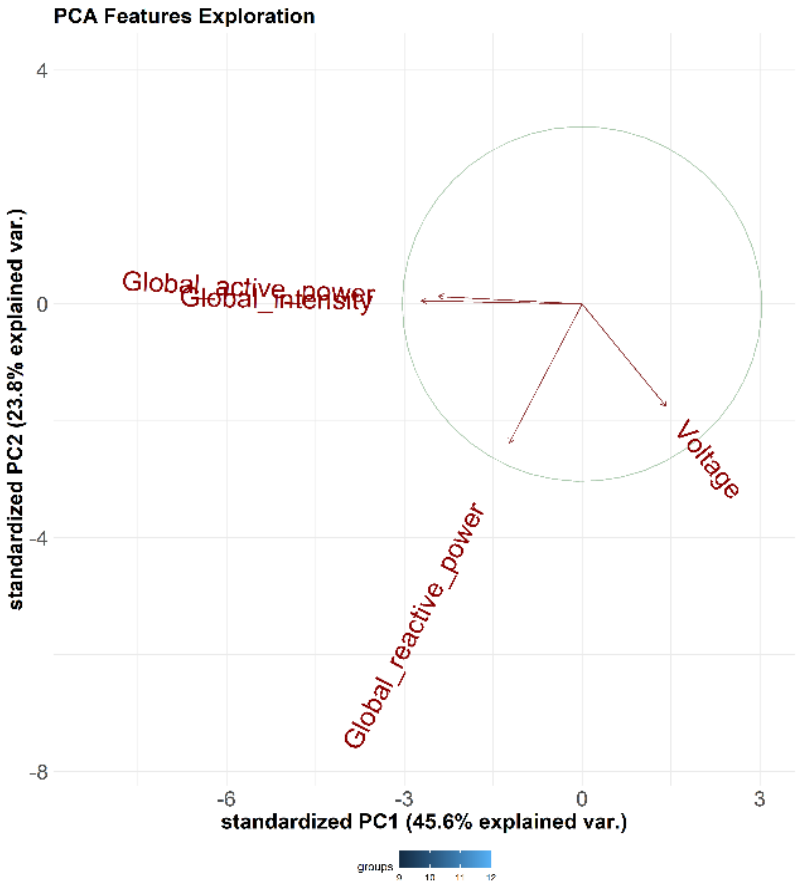
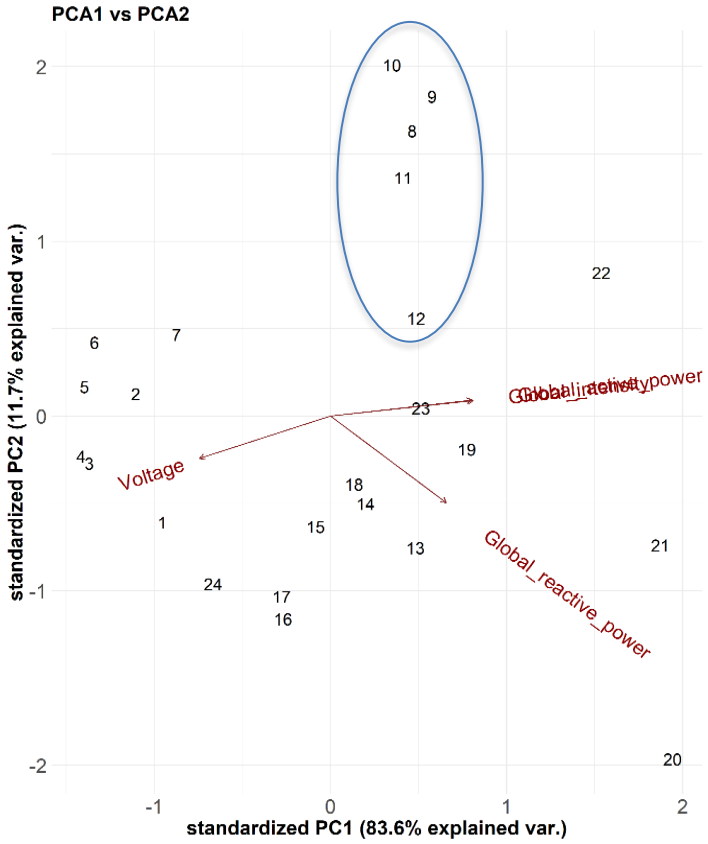


Figure 12 – PCA analysis for the raw data set for time window 9am-1pm (aggregated for all years)

Figure 11 demonstrates the difficulty in visualizing the variance of distributions in PCs. To get better results for the analysis, the data will be aggregated annually (by hours) and the PCA will be visualized.

Based on the hourly aggregation of data for all 4 years, PC1 corresponds to 83.6% of the total variance while PC2 corresponds to 11.7%. By knowing the position of a sample value with respect to PC1 and PC2, it is rather simple to get an accurate view on where the value stands with respect to other samples, as PC1 and PC2 account for approximately 95% of the variance. The following several graphs and tables demonstrate further explorations of the dependent features.

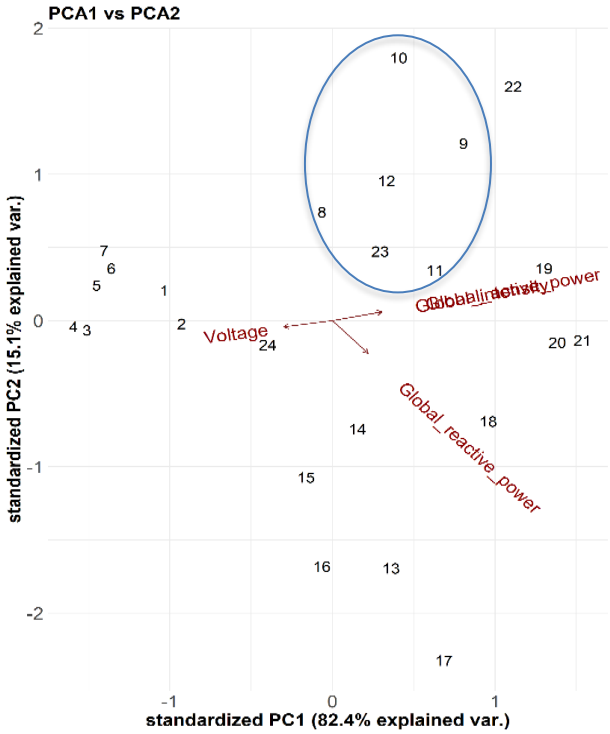



```
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.8284 0.6849 0.42936 0.06121
Proportion of Variance 0.8357 0.1173 0.04609 0.00094
Cumulative Proportion  0.8357 0.9530 0.99906 1.00000
```

Figure 13 – PCA analysis for the complete dataset (hourly aggregation of data for all 4 years)

Figure 13 shows the distribution between PCA1 and PCA2 and makes it easy to visualize which hours belong to which group. The Global Active Power and Global Intensity features are overlapping which means that they have a high correlation and dependency. Conversely, Voltage and Global Reactive Power are separate from other features. Hence, for the multivariate HMM training, **Global Active Power** and **Global Intensity** will be selected as the dependent variables.

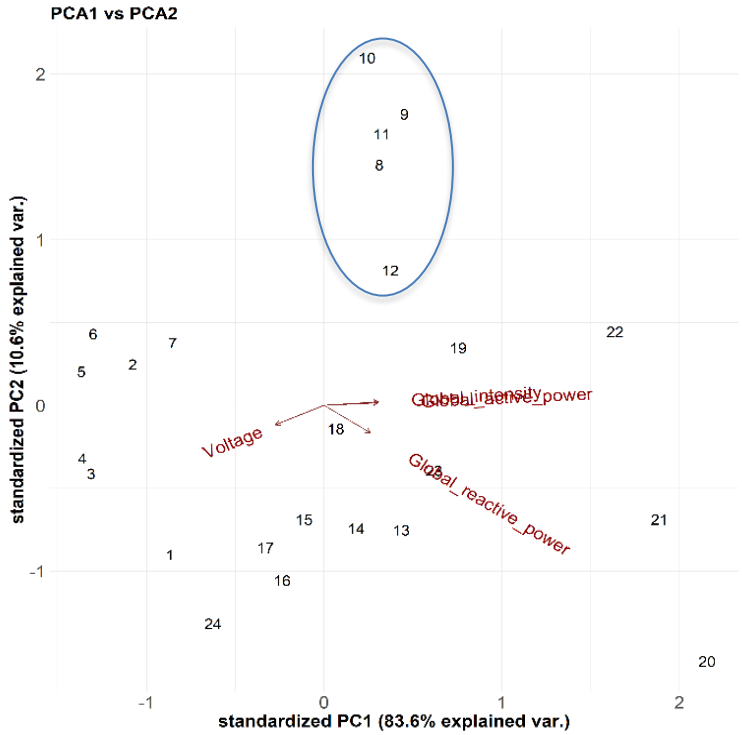
Figure 13 also shows how daytime hours—9am to 1pm, in particular—are closely related to each other and are separated from rest of the day. Similar to the methodology used for data exploration, in order to further analyse and support the feature selection, individual years are analysed for PC distribution. The following several graphs summarize the findings for PCA analysis year-by-year.




```
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.8160 0.7762 0.30932 0.06449
Proportion of Variance 0.8244 0.1506 0.02392 0.00104
Cumulative Proportion  0.8244 0.9750 0.99896 1.00000
```

Figure 14 – Results of PCA analysis (year 2006)

Figure 14 shows the PCA analysis on hourly aggregated data for year 2006. It includes the distribution of hours and follows the same pattern as per graph for all years data in **Figure 13**. Global Active Power and Global Intensity are overlapping, which supports our conclusion that they are correlated and dependent. In addition, among daytime hours (9am to 1pm), features are closely related to each other. No further analysis is required for PC3 and PC4 as PC1 and PC2 account for 97% of the variance. The analysis confirms that Global Active Power and Global Intensity are highly correlated and dependent.

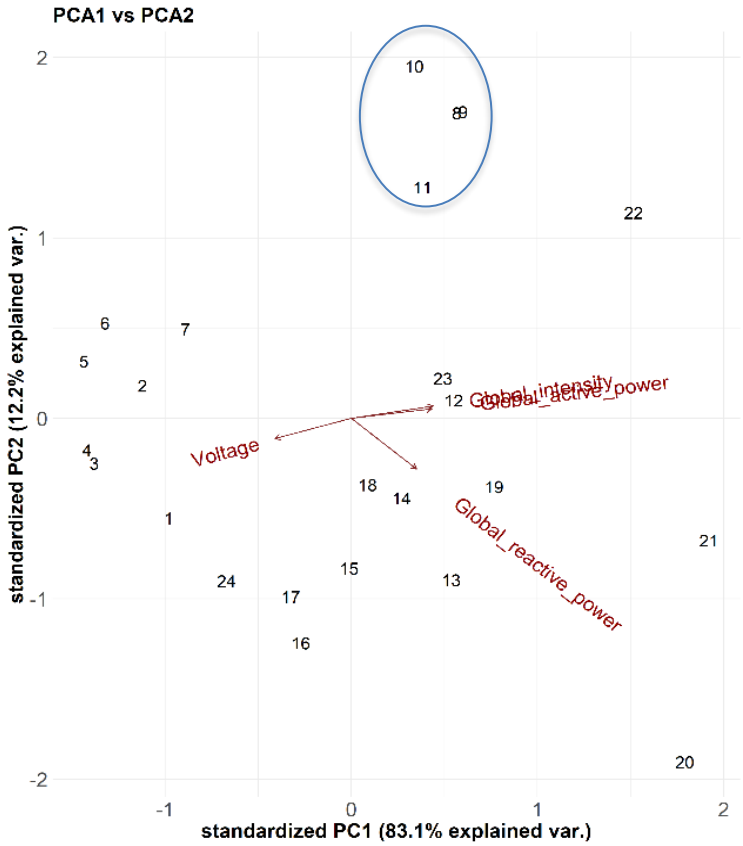



```
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.8291 0.6522 0.47001 0.08931
Proportion of Variance 0.8364 0.1063 0.05523 0.00199
Cumulative Proportion  0.8364 0.9428 0.99801 1.00000
```

Figure 15 - Results of PCA analysis (year 2007)

Figure 15 shows that the PCA analysis on hourly aggregated data for year 2007 follows the same pattern for correlation and dependency for Global active power and Global Intensity. As well, it produces a similar pattern for daytime hours (9am-1pm) as the previous

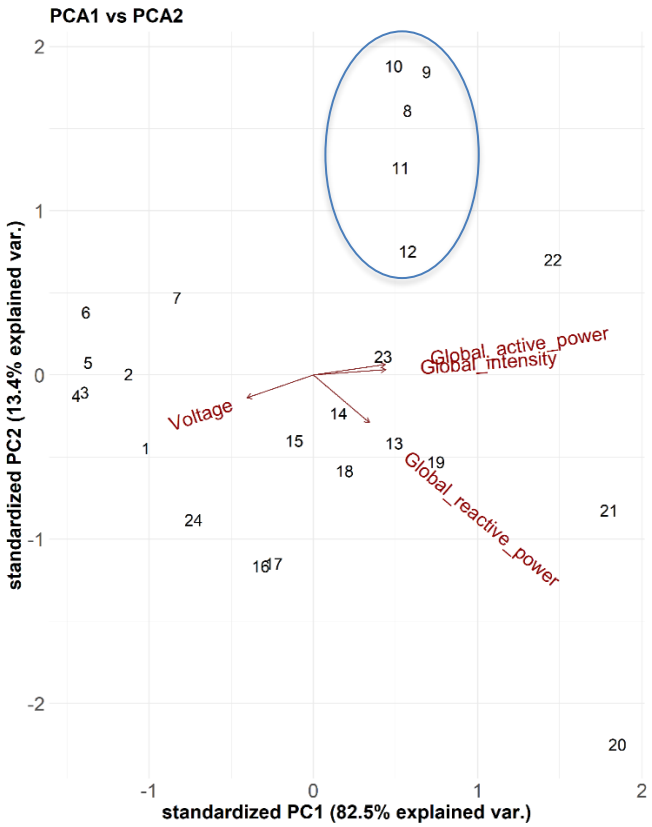
graphs. No further analysis is required for PC3 and PC4 as PC1 and PC2 account for 94% of the variance. The analysis confirms the conclusion of correlation and dependency.




```
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.8236 0.6985 0.42636 0.07038
Proportion of Variance 0.8314 0.1220 0.04545 0.00124
Cumulative Proportion  0.8314 0.9533 0.99876 1.00000
```

Figure 16 - Results of PCA analysis (year 2008)

Figure 16 shows the PCA analysis on hourly aggregated data for year 2008. The summary of components shows PC1 and PC2 account for 96% of variance. The graph displays the same pattern of correlation and dependency for Global Active Power and Global Intensity. As well, Voltage and Global Reactive Power do not correlate with any features. Interestingly, the daytime hours that display similar patterns are from 8am to noon as opposed to 9am-1pm.




```
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.8163 0.7331 0.40044 0.05628
Proportion of Variance 0.8248 0.1343 0.04009 0.00079
Cumulative Proportion  0.8248 0.9591 0.99921 1.00000
```

Figure 17 - Results of PCA analysis (year 2009)

Figure 17 shows the PCA analysis on hourly aggregated data for the final year 2009. It displays follows the same continuous pattern of correlation and dependency for Global Active power and Global Intensity and the summary of components shows PC1 and PC2 account for 96% of variance. The daytime hours of 8am-1pm display similar patterns and belong to one subset. As such, the data set for the last year reconfirms our conclusion that Global Active Power and Global Intensity are correlated and dependent.

As global active power and global intensity show a strong continuous correlation and dependence throughout the PCA analysis of the data over 4 years, they are used in training the multivariate HMM models. As well, in annual PCA analysis, the pattern of variables within the selected daytime window of 9am to 1pm lies in the same subset, hence, the choice is affirmed for further analysis of training and testing of HMMs and anomaly detection.

3.3. TRAINING AND TESTING OF HIDDEN MARKOV MODELS

For the different HMMs, the same time frame of 9am-1pm for weekdays and weekends is used throughout the training process. The given 'training' data set is initially divided into two parts: the timeframe of December 16th, 2006 to April 30th, 2009 is selected as the subset used to train HMMs whilst the rest of the data (from May 1st, 2009 to December 1st, 2009) is used to test the trained models. This produces 124 weekdays/weekend days ($\frac{4}{5}$ of the data size) for the training data set and 31 days ($\frac{1}{5}$ of the data size) to the test data set. To choose the most optimal state, HMMs are trained and with the number of hidden states ranging from 5-15 (for univariate weekdays models) and 5-22 (for univariate weekends models), then the results are plotted to perceive any patterns in log-likelihood and BIC values corresponding to each number of states. Next, the most optimal state is selected based on the largest log likelihood closest to zero and the lowest BIC value. If the log-likelihood at the current and subsequent states begins to exceed zero, one assumes that such models are unacceptable and can be ignored as likelihood values are between zero and one, hence their log values must be negative.

For weekdays, the most optimal univariate model is determined early at 7 states, because the log-likelihood is at its maximum (closest to zero). After 7 states, the log-likelihood

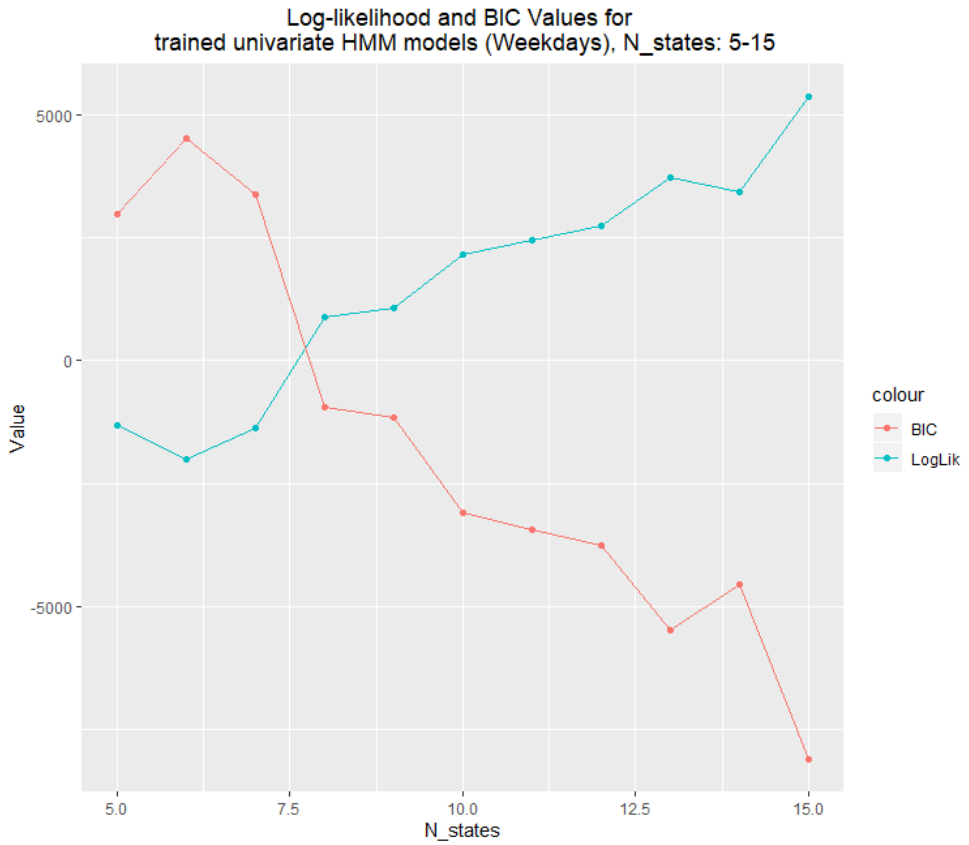


Figure 18 – Log-likelihood and BIC values for trained univariate weekdays HMM model, number of states range from 5-15.

begins to enter the positive region. This ultimately determines that the number of states for the univariate weekdays HMM model is deemed to be sub-optimal after 7 states.

For the second model, weekend data is analysed and trained for HMM models. Based on the plot in **Figure 19**, the growth of log-likelihood values and decay of BIC values are gradual as the number of states increase and the graph appears to flatten for both values after training for ~19 states. This implies that additional states do not result in further increased performance of the subsequent HMM models. As such, the selection of the number of states is based on finding the minimum BIC value, which is obtained when number of states is 19 (log-likelihood values for states 19, 21 and 22 are similar). Hence, the HMM model with **19 states** is selected as the best univariate HMM model for the weekend training data.

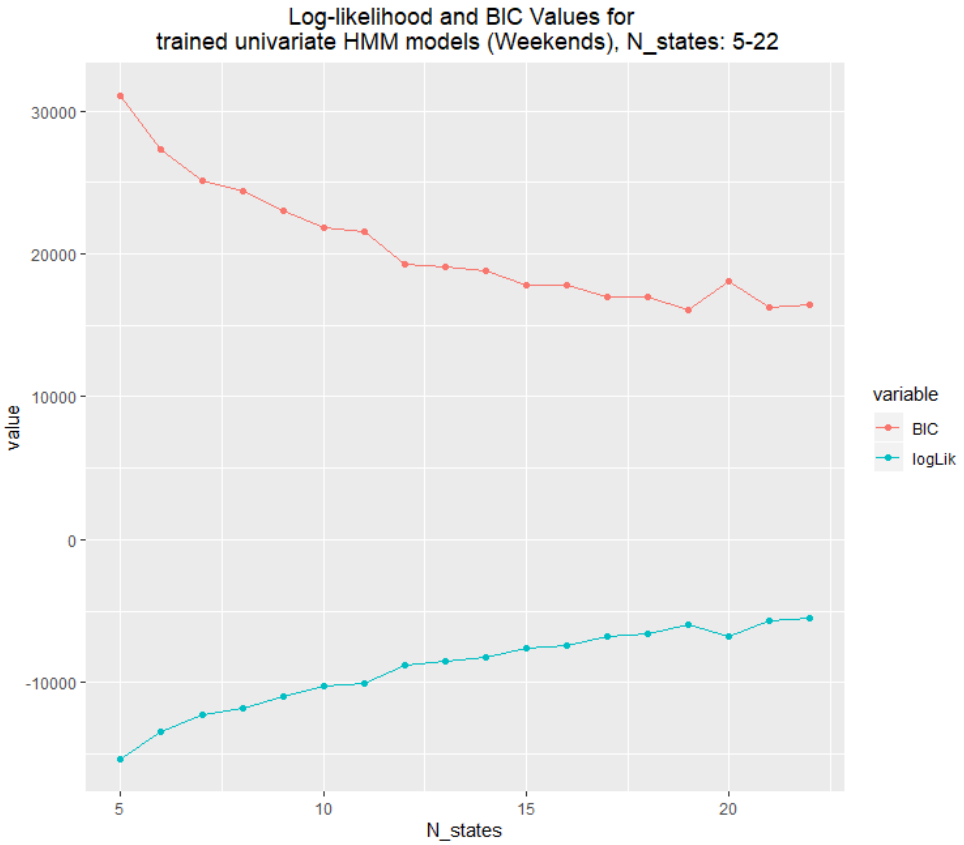


Figure 19 - Log-likelihood and BIC values for trained univariate weekends HMM model, number of states range from 5-22.

Expecting that the number of states to correspond to the number of activities in households involving electricity consumption, it is anticipated that the optimal HMM model for the weekend data to be more complex and to have more states than the optimal model for the weekdays data. The variety and length of activities over the weekends is less deterministic—they could range from cooking to watching tv, playing video games, etc.—whereas on weekdays, between the hours of 9am

and 1pm, the majority of activities are expected to be outdoors at work or school. Hence, the number of states for univariate HMM models are reasonably justified as stated.

For the multivariate models, to assist with the selection of rightly fitted HMMs, after training for states 5-25 for both weekday and weekend data, the differences between training and testing log-likelihood values are calculated and compared for states 21-25. This is done to avoid overfitting the data with complex HMMs as well as the fact that the training log-likelihood values are similar. Based on the analysis, a multivariate model with 22 states is selected for both weekdays and weekends.

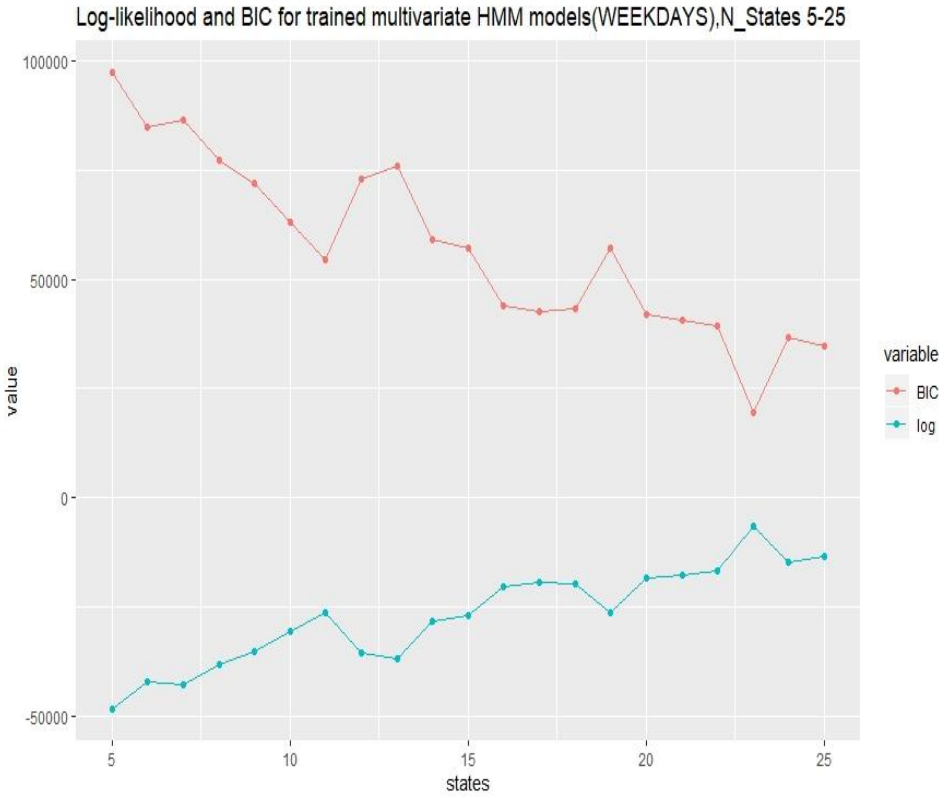


Figure 20 - Log-likelihood and BIC values for trained multivariate weekdays HMM model, number of states range from 5-25.

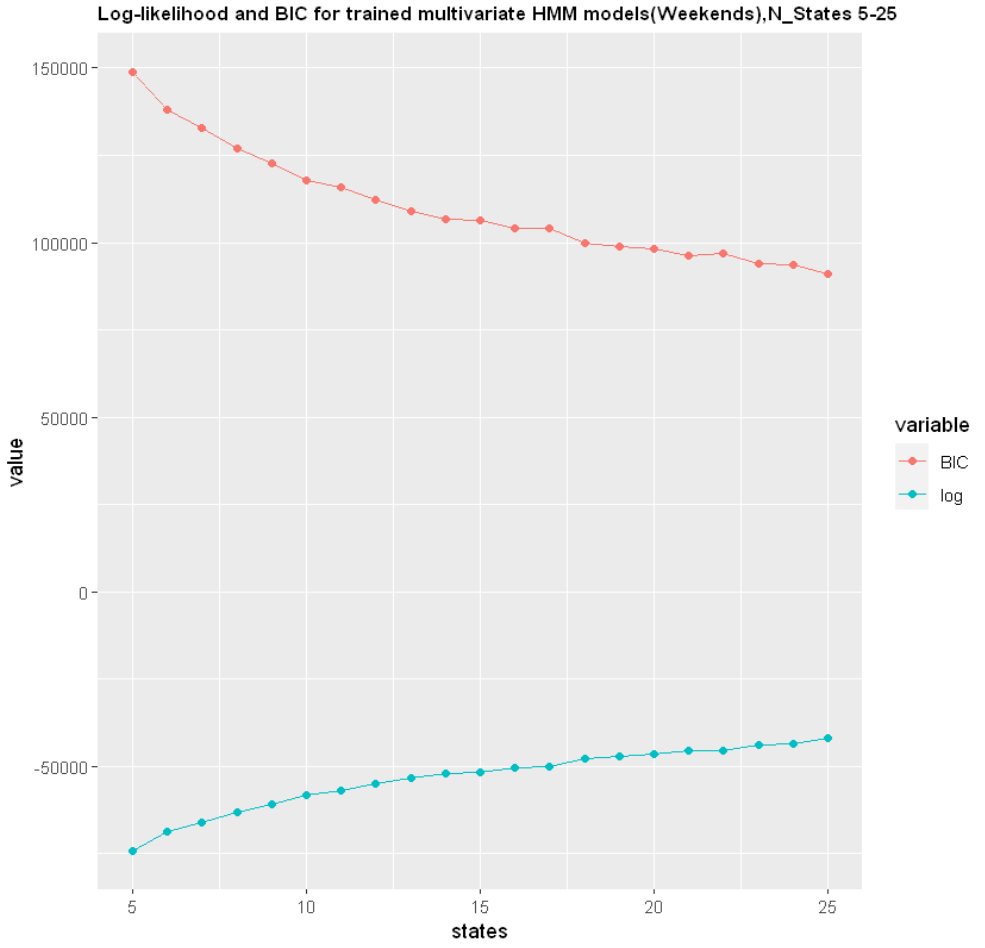


Figure 21 - Log-likelihood and BIC values for trained multivariate weekends HMM model, number of states range from 5-25.

A summary of the essential model characteristics for most optimal HMM models is included in *Experimental Results* (**Table 3**) below.

Table 3 – a summary of *Experimental results* obtained from training and testing univariate and multivariate HMM models for weekdays and weekends data

Characteristics of Training HMMs (Date Range of Data: 12-16-2006 to 04-30-2009)				
Dependent Variable(s)		**N-states**	**log-likelihood**	**BIC**
Univariate (Weekdays)				
global_active power		**7**	**-1368**	**3374**
Univariate (Weekends)				
global_active power		**19**	**-5977**	**16053**
Multivariate (Weekdays)				
global_active_power	**global_intensity**	**22**	**-16665**	**39207**
Multivariate (Weekends)				
global_active_power	**global_intensity**	**22**	**-44471**	**94819**
Characteristics of Test HMMs (Date Range of Data: 05-01-2009 to 12-01-2009)				
Dependent Variable(s)		**N-states**	**log-likelihood (test)**	**log-likelihood/4 (training)**
Univariate (Weekdays)				
global_active_power		**7**	**337**	$\frac{1}{4} \times -1368 = -342$
Univariate (Weekends)				
global_active_power		**19**	**-1049**	$\frac{1}{4} \times -5977 \approx -1494$
Multivariate (Weekdays)				
global_active_power	**global_intensity**	**22**	**-4973**	$\frac{1}{4} \times -16665 \approx -4166$
Multivariate (Weekends)				
global_active_power	**global_intensity**	**22**	**-9769**	$\frac{1}{4} \times -44471 \approx -11118$

For all the comparisons between the training and test log-likelihoods, the quality of the model is determined by the difference between log-likelihoods: if the difference between the values is smaller than 1000, the model is considered accurate to be used for further analysis of detecting contextual anomalies. Furthermore, considering that the training of HMMs is an abstract process—with limitations imposed by the forward-backward algorithms used in "depmixS4" package in R, —there is a chance that the training of HMMs could be misconstrued. Therefore, a difference of 1000 in training and testing log-likelihood values is determined to be a reasonable modelling error.

3.4. ANOMALY DETECTION

For point anomaly detection three different pairs of thresholds are defined for weekends and weekdays separately. The range of threshold values are [-0.75, 0.75], [-1.5,1.5], [-2.25,2.25] for weekdays and [-1.25, 1.25], [-2.5,2.5], [-3.75,3.75] for weekends. The rationale for choosing these

ranges is based on the observation of significant jump in electrical consumption behaviour within the 5 provided data sets. Based on these ranges, 2×three sets of outliers would be obtained for each set. These three sets can be interpreted from two different aspects: a) the amount of electricity savings and/or wastages (e.g. the first threshold range would be for household with the highest savings in electricity consumption), b) the severity of system intrusion and the amount of abnormal behaviour in electrical consumption cause by hackers and intruders. Based on the moving average analysis and the observed behaviour of the five data sets, it is generally understood that the first three show more normal behaviour in comparison to the data sets 4 and 5 on a typical weekday (Monday, January 25th, 2010) and a typical weekend day (Sunday, January 24th, 2010) between the hours of 9am-1pm.

Figure 22 shows one sample result for point anomaly detection analysis using the moving average technique. The rest of the 29 graphs can be found within the zip file, in the "Point_ Anomaly_ Detection_ Results" directory.

Figure 22 - Moving Average analysis results for weekday, 9am-1pm (test data #1, threshold value: ±0.75)

The results for contextual anomaly detection, comparing the log-likelihood of trained HMM models with those of the provided 'anomalous' test data, is summarized in the table below. For each of the 5 provided data sets, the log-likelihood values are calculated and tabulated for the same time windows over weekdays and weekends. Note that as the test sets provide power usage data from December 1st, 2009 to November 26th, 2010 (resulting in 51 weekdays and 52 weekend days), the log likelihoods of trained models are multiplied by $\frac{51}{124}$ and $\frac{52}{124}$ to make the results comparable.

Table 4 – Summary of contextual anomaly detection results

Contextual Anomaly Detection Results—Comparison of log-likelihoods	
log-likelihood (test)	**log-likelihood (training)**
Univariate (Weekdays) – N-states of optimal HMM model: 7	
Test data set #1	
-39343	$\frac{51}{124} \times -1368 \approx$**-563**
Test data set #2	
-40020	$\frac{51}{124} \times -1368 \approx$**-563**
Test data set #3	
-39343	$\frac{51}{124} \times -1368 \approx$**-563**
Test data set #4	
-194484	$\frac{51}{124} \times -1368 \approx$**-563**
Test data set #5	
-193989	$\frac{51}{124} \times -1368 \approx$**-563**
Univariate (Weekends) – N-states of optimal HMM model: 19	
Test data set #1	
-18655	$\frac{52}{124} \times -5977 \approx$**-2506**
Test data set #2	
-18931	$\frac{52}{124} \times -5977 \approx$**-2506**
Test data set #3	
-18655	$\frac{52}{124} \times -5977 \approx$**-2506**
Test data set #4	
-78868	$\frac{52}{124} \times -5977 \approx$**-2506**
Test data set #5	
-79586	$\frac{52}{124} \times -5977 \approx$**-2506**
Multivariate (Weekdays) – N-states of optimal HMM model: 22	
Test data set #1	
-12416.12	$\frac{51}{124} \times -16665 \approx$**-6854**
Test data set #2	
-12347.53	$\frac{51}{124} \times -16665 \approx$**-6854**
Test data set #3	
-12416.12	$\frac{51}{124} \times -16665 \approx$**-6854**
Test data set #4	
-70578.85	$\frac{51}{124} \times -16665 \approx$**-6854**
Test data set #5	
- 69816.45	$\frac{51}{124} \times -16665 \approx$**-6854**
Multivariate (Weekends) – N-states of optimal HMM model: 22	

Test data set #1	
-12014.00	$\frac{52}{124} \times -44471 \approx$**-18649**
Test data set #2	
-12034.97	$\frac{52}{124} \times -44471 \approx$**-18649**
Test data set #3	
-12014.00	$\frac{52}{124} \times -44471 \approx$**-18649**
Test data set #4	
-40935.09	$\frac{52}{124} \times -44471 \approx$**-18649**
Test data set #5	
-40869.22	$\frac{52}{124} \times -44471 \approx$**-18649**

Based on the results in **Table 4**, it is apparent to see that the training data fails to match the data provide according to the major differences in log-likelihood values between the training models and the 'test' models. When the univariate training models for weekday data and weekend are tested using the 5 data sets, the testing log-likelihood is significantly smaller than the training log-likelihoods. Similarly, when the multivariate training models are tested, they displayed similar results. Multiple factors that could have influenced these discrepancies include the fact that the test data contains more injected anomalies than the training data and deviate from 'normal' behaviour in energy consumption, or the fact that R is computing the results inaccurately due to multiple running functions and hidden variables in the background that are influencing the testing algorithms. Nevertheless, there is a clear pattern that has surfaced during the testing phase across all HMMs. In the results, all the log-log-likelihoods in test data sets #1, #2, and #3 all return similar values and are closer to the log-likelihoods of the model that was trained in comparison to test data sets #4 and #5. Whereas, the test HMMs for test data #4 and #5 give much smaller log-likelihoods with values that are drastically different from the training log-likelihoods. These results support the fact that test data #4 and #5 are much more anomalous than test data #1, #2 and #3. Therefore, it can be inferred that there are many injected anomalies that reside within test data #4 and #5.

4.0. PROJECT CHALLENGES

There are several challenges in the process of analysing the data for behavioural-based intrusion detection. One main such challenge is the training of univariate and multivariate models. This is heavily due to the extensive amount of time required to train HMM models with many states to determine and select the most optimal model while avoiding complexity and overfitting the training data set. Furthermore, training multiple Markov models requires much of CPU processing power on an individual's laptop or computer and limits the number of concurrent tasks on each PC. Hence, the task of training the HMMs were delegated among all the team members. However, all computers train and process HMMs at different rates and return dissimilar models for the same time windows and the same number of dependent variables. Therefore, in addition to the limitations of algorithms in HMM packages used in R, the model training is highly dependent on the hardware features of the system used. As such, another challenge in the project is the process of determining the accuracy and legitimacy of one team member's results from model training and testing in comparison to others. Considering that the log-likelihoods are the major selection factor when one selects the optimal HMM models as well as determining anomalous behaviour of the test data, the process of training the HMMs have to be executed meticulously and perhaps performed multiple times on the same PC. However, due to unknown number of hidden states as well as the background variables in R, it is difficult to determine whether HMMs are exhibiting random characteristic where the most optimal number of states is a small value or rather a large number. This means that one group member can have a complete training model and test result due to their optimal HMMs requiring a small number of states but be waiting on other team members with higher number of states to find their most optimal model. This causes an interruption in both communication and progress as the comparison of each team member's log-likelihood is essential to perform anomaly detection tasks as well as ensuring the models are not overcomplicated/overfitted. On the topic of communication, one major challenge that affected the team is the unorthodox distance

communication due to COVID-19 and the provincial quarantine. Teamwork and consistent in-person meetings are an essential step for a fluid execution of the project. Overall, all project challenges served as great obstacles but are now valuable lessons learned that all team members can utilize moving forward.

5.0. LESSONS LEARNED

The project is a valuable opportunity for the team to encounter and analyse a real-world data science problem using available data sets. Some of the major lessons that learned through the course of completing this project are summarized below.

Firstly, analysing data sets without learning the proper terminologies and understanding the associated field of knowledge and cybersecurity concepts is simply impossible. Referencing the course material and performing external research to understand certain concepts in highly beneficial. Understanding and applying statistical analysis on data is one area that will be cherished, considering that the knowledge of these subjects is highly invaluable and sought after in many careers.

In addition, improvement of time management, communication and teamworking skills are some of the other outcomes. Regular weekly/daily scrum meetings enables each team member to have a meticulous mindset and a well-structured and concise workflow for completion of the project.

Furthermore, the training and testing process of HMMs are practiced and better understood. For point anomaly detection, defining an adequate threshold is difficult without the knowledge of notion or an expert advice on the definition of boundaries on normal behaviour of electrical consumption. Often, defining regions that show normal behaviour is not precise. Contributing reasons include the prevalence of noise and abnormal behaviour of the data set due to injected anomalies. To better determine the normal behaviour of the data and better thresholds for point anomaly detection as well as contextual anomalies, it may be beneficial to create multiple versions of the HMMs for the same day and time window (on different machines) or perhaps look at multiple time windows to get a better sense of normal behaviour for the selected dependent variables.

Finally, the importance of data exploration is learned as large data sets need to be analysed using multiple techniques to derive the best observable patterns. For this purpose, PCA analysis and calculation of correlation matrices are employed to unravel the underlying dependencies between different parameters of the data sets. In general, it is important to note that more data does not necessarily lead to better performance of models and results or more profound insights.

6.0. CONCLUSION

Through the project, the importance of the risks involved within power grid cybersecurity is outlined. As well, one modern solution to prevent and mitigate the possible impacts of electricity grid disruption is addressed and analysed in detail: the point and contextual anomaly detection using multiple hidden Markov Models is critical to prevent cybercrimes on electrical power grids at this day and age. Despite the tedious processes for data exploration and high use of resources for training and testing HMM models, the techniques being applied and demonstrated in this project are some of the only means of cybersecurity tools for defence.

7.0. ACKNOWLEDGEMENTS

All team members had equal contribution to submission of the project report and implementation of the project code.

8.0. REFERENCES

Electric utility responses to grid security issues. (2006). Retrieved from IEEE Xplore Digital Library: https://ieeexplore.ieee.org/document/1597993

National Electric Grid Security and Resilience Action Plan. (2016). Retrieved from Natural Resources Canada: https://www.nrcan.gc.ca/sites/www.nrcan.gc.ca/files/energy/pdf/Canadian%20Action%20Plan_EN.PDF

Russian operation hacked a Vermont utility, showing risk to U.S. electrical grid security, officials say. (2016, December 31). *The Washington Post*. Retrieved from https://www.washingtonpost.com/world/national-security/russian-hackers-penetrated-us-electricity-grid-through-a-utility-in-vermont/2016/12/30/8fc90cc4-ceec-11e6-b8a2-8c2a61b0436f_story.html