CMPT 318 Cybersecurity

Anomaly Detection of Electric Power Grids Using Hidden Markov Models

Issued

Fall 2019

Members of Group 11

Amanda Pak 301378675 apak@sfu.ca

Christopher Lam 301303636 cla318@sfu.ca

James McTaggart 301079987 jmctagga@sfu.ca

Robin Truong 301053650 rlt5@sfu.ca

# Table of Contents

[**Table of Contents**](#_izzpm3hcxn20) **1**

[**Variable Definitions**](#_h89e02e3a7cw) **2**

[**Phase 1. General Data Exploration**](#_7fqcbz9vg5az) **4**

[**Anomaly Detection**](#_f9tqfghrwu8b) **13**

[**Phase 2. Single Feature HMMs**](#_8rni49dwy7hp) **18**

[**Log-Likelihood**](#_hw4t68jd2u8q) **18**

[**Bayesian Information Criterion (BIC)**](#_b6skg5ae0ti1) **19**

[**Time Period Selection**](#_e5kqaj2us3m5) **19**

[**Cross Validation**](#_28hrr4n8zyka) **26**

[**Three Feature HMM Testing**](#_h86g8xjkewvw) **27**

[**HMM Summary**](#_8b4jhobsosxm) **31**

[**Phase 3. Anomaly Detection Approach**](#_1pfydrokizi2) **32**

[**Questions**](#_9voc7t7okia0) **35**

[**Appendix A - Charts and Tables**](#_bbuapzpejdcp) **38**

[**Appendix B - Tanvir’s Individual Work**](#_jfhpu56pmzu6) **50**

[**Term Project.docx**](#_ccuanj8kddyf) **50**

[**Term Project.R**](#_xl1tuv2d5lmf) **59**

[**Improvement of ass1.R**](#_c4jr2ubkixop) **66**

# 

# Variable Definitions

A household uses electrical energy to run devices. The electricity is monitored by several power, and voltage meters. The data collected is mainly used for billing the clients but can also be used for anomaly detection. The provided dataset contained the following variables:

**Global\_active\_power**: The household global minute-averaged active power (in kilowatts kW). When a household uses electricity, the active power meter, connected to the main feed into the house, records the total active power draw. This power measurement is used by the electrical utility company to bill the client on their electrical energy consumption.

**Global\_reactive\_power**: The household global minute-averaged reactive power (in kiloVars). Reactive power is defined as the where S is the apparent power (VI). Reactive power is generally generated by motors and transformers or other inductive devices. These devices cause a phase shift in the current. Electrical utility companies do not really use this measurement for anything other than monitoring the amount of Vars injected into their system. Clients will not get charged on the amount of Vars being used but if users inject too much reactive power into the system, the utility may charge a fine for too much reactive power. A way to mitigate reactive power injection to the grid is to use power factor correction (PFC) units.

**Voltage:** The minute-averaged voltage (in volts). This voltage is the grid voltage that feeds into the household. The higher the load on the system, the lower the voltage.

**Global\_intensity**: The household global minute-averaged current intensity (in amperes). The higher active+reactive power consumed, the higher the current.

**Sub\_metering\_1:** The power sub-metering No. 1 (in kiloWatts). It corresponds to the kitchen, containing mainly a dishwasher, an oven and a microwave (hot plates are not electric but gas powered).

**Sub\_metering\_2:** The power sub-metering No. 2 (in kiloWatts). It corresponds to the laundry room, containing a washing-machine, a tumble-drier, a refrigerator and a light.

**Sub\_metering\_3**: The power sub-metering No. 3 (in kiloWatts). It corresponds to an electric water-heater and an air-conditioner.

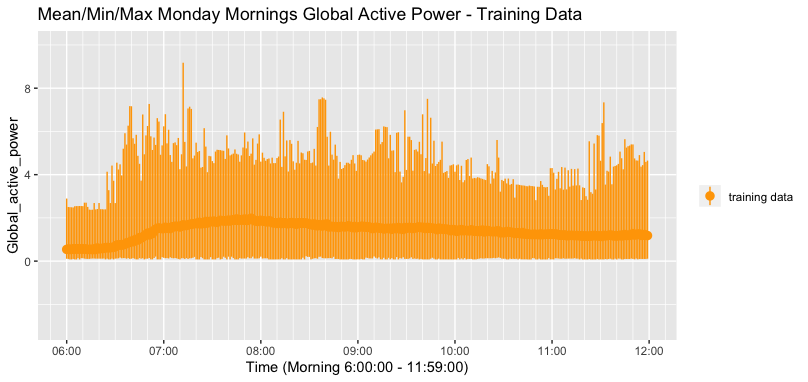
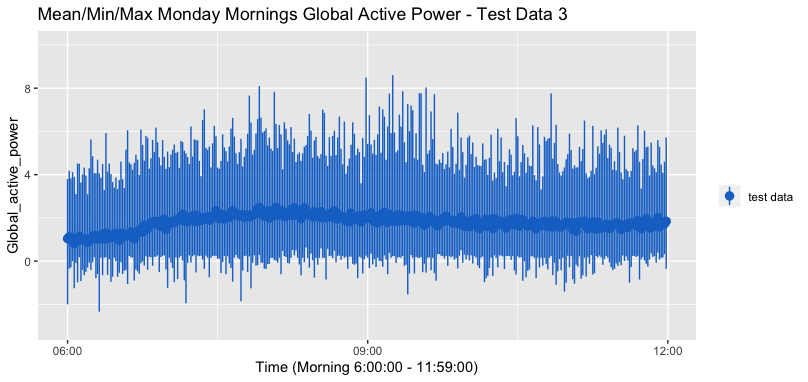
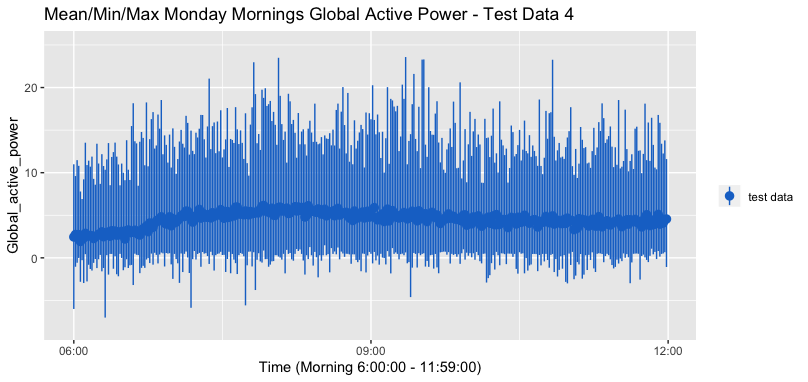
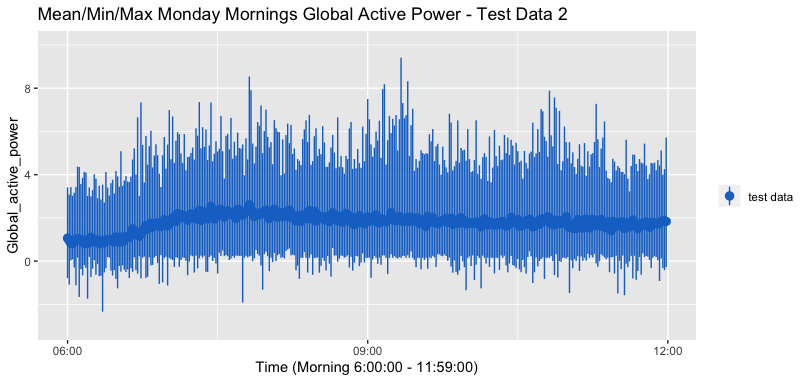
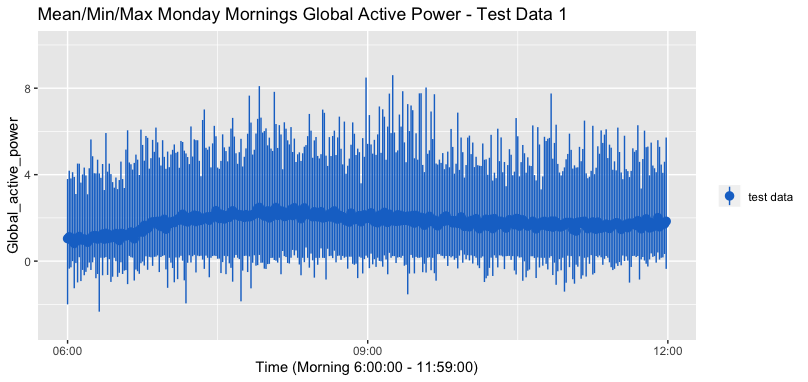
# 

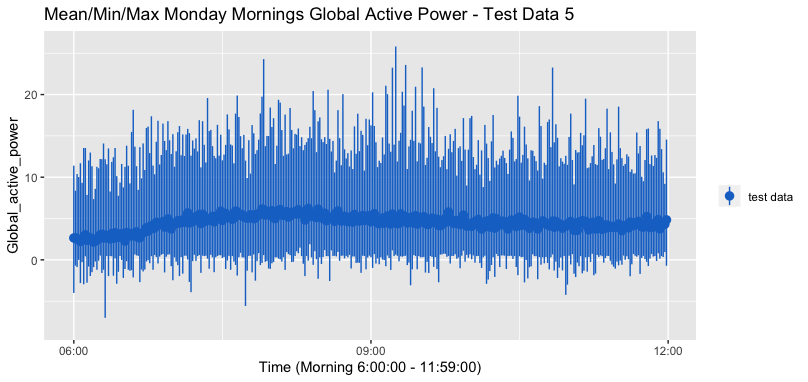
# Phase 1. General Data Exploration

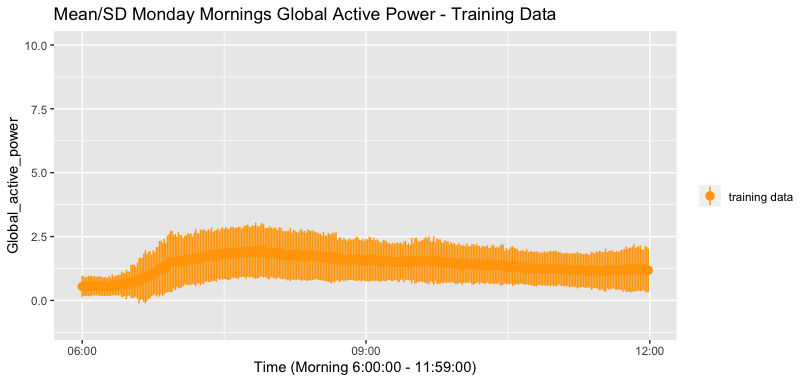
**Summary Statistics**

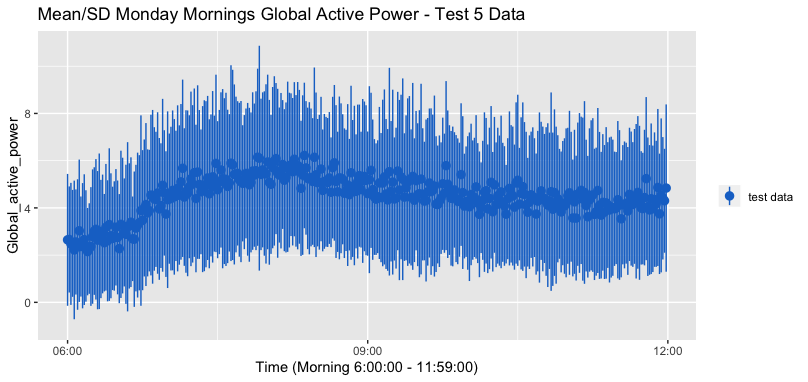
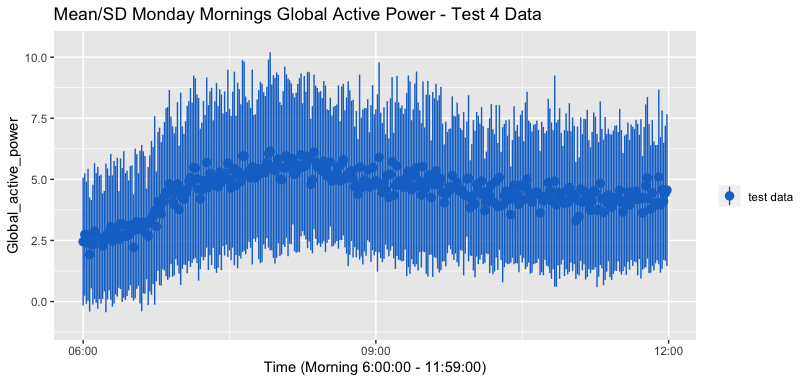
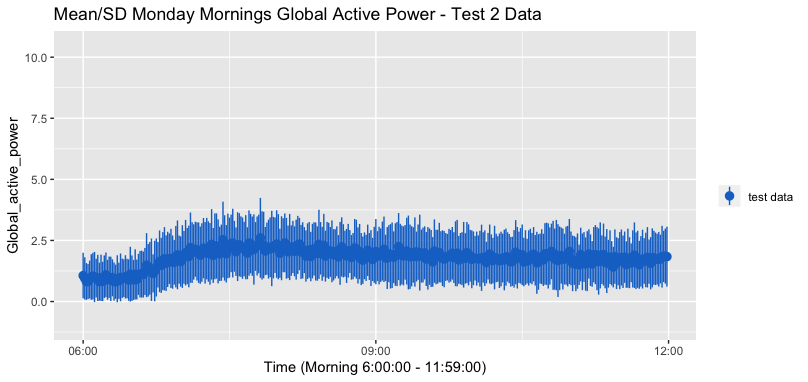
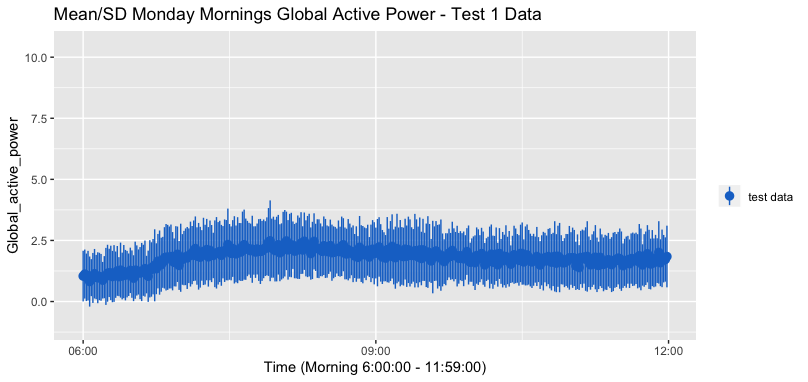
The mean, maximum, minimum, and standard deviation for global active power in a window of Monday mornings (6:00 am - 11:59 am) was compared between the training data and the testing data. The results show that all training data measurements are positive, while the test data sets have negative values at various minutes in the window. Test data sets 1, 2, and 3 shared a similar range to the training data, while tests 4 and 5 had much higher maximums. Overall, the trend of test data 4 and 5 had a much higher mean compared to the training data. The standard deviation of global active power for Monday mornings was examined and showed the training data, test 1, test 2, and test 3, had a similar dispersion in data but test 4 and test 5 have a much larger dispersion.

The testing data represents normal behaviour for global active power. The statistics of all data sets suggest that tests 1, 2, and 3 are relatively close to normal while tests 4 and 5 deviate significantly from the observed normal behaviour of global active power for Monday mornings.







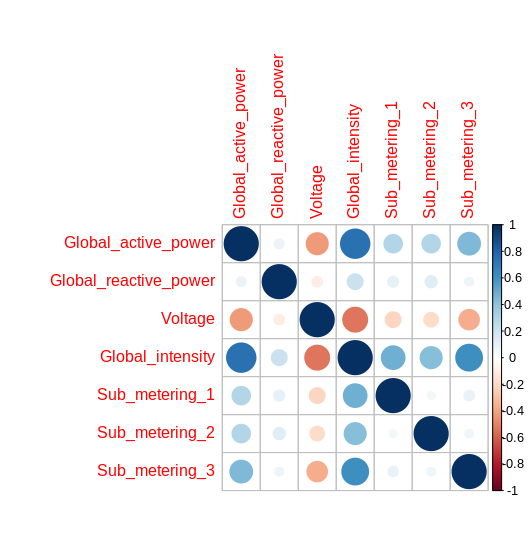


**Feature Correlation**

Before we jump into the details it is worth establishing a few of the basics for those unfamiliar with data analysis. Firstly, we are about to examine multivariate time series which are a collection of univariate time series. Essentially a univariate time series is a collection of data points over a period of time. For example, if you were to track the temperature in your backyard over the course of a year at regular intervals you would have created a univariate time series. Now a multivariate time series just takes this one step further and it includes more than one data point over the same time interval. In our case we have a multivariate time series that includes the following variables:

Global\_active\_power, Global\_reactive\_power, Voltage, Global\_intensity, Voltage, Global\_intensity, Sub\_metering\_1, Sub\_metering\_2, Sub\_metering\_3

Now we are presented with an overwhelming amount of information so it is worth examining how we can extract useful bits of information from the vast quantity of data. That is where correlation coefficients become useful. Correlation coefficients are a measure of the strength of the association among the various variables. In our case the correlation coefficient is this:



Correlation Matrix

The correlation matrix is used to confirm our understanding of the given parameters. We know that each sub\_meter measures power usage of different circuits in the house therefore the sub\_meters should be independent of each other and show a low correlation coefficient with each other. If the majority of power usage is active power, we know that global intensity (current) will have a high correlation coefficient with global active power. Global intensity would be made up of all the current going through each sub\_meter circuit therefore each sub\_meter should have a high correlation ratio with the global\_intensity and global\_active\_power.

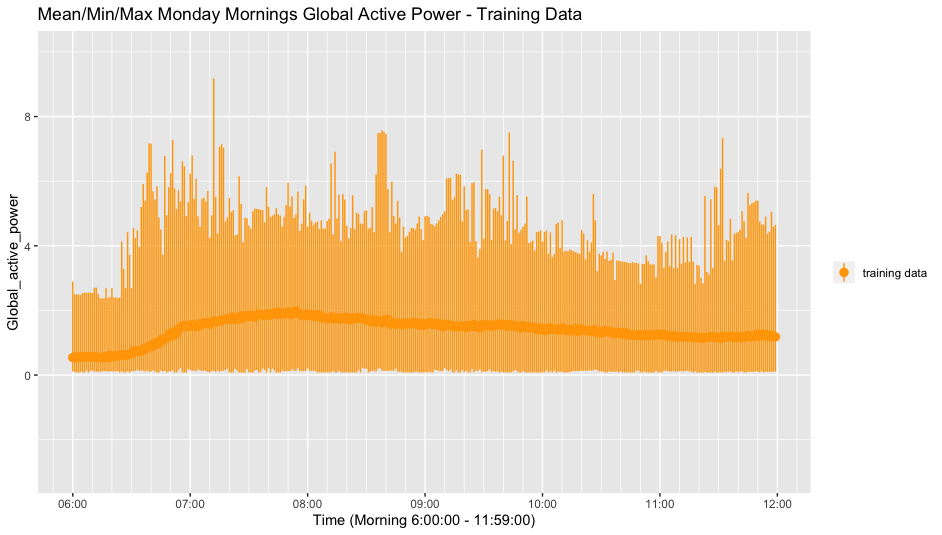
The power output from the utility provider is constant and is equal to: P=VI. When the current draw increases (increased global\_intensity), the voltage decreases. This should mean that there is a negative correlation coefficient between Voltage and Current. From the correlation coefficient matrix, we observe that voltage and Global\_intensity have a negative correlation which confirms our intuition. The chart also indicates that certain data points, Global\_reactive\_power for example, aren’t particularly correlated positively or negatively at all with any of the other features. This is most likely due to the low amount of reactive power being used by common households because they generally don’t power large motors or transformers (inductive loads).

If we were to include all of the features in our comparison the information would not be particularly meaningful because many of the features are not correlated very highly. It would be an overload of uncorrelated information. However, if we were to select certain features that had reasonably high correlation and then we were to analyze those data sets together we might be able to form certain inferences about the relationship between the data. Ultimately, we just need to limit the number of features that we are examining together at one time because these particular features are a mix of correlations. The data is much more manageable in smaller chunks.

## 

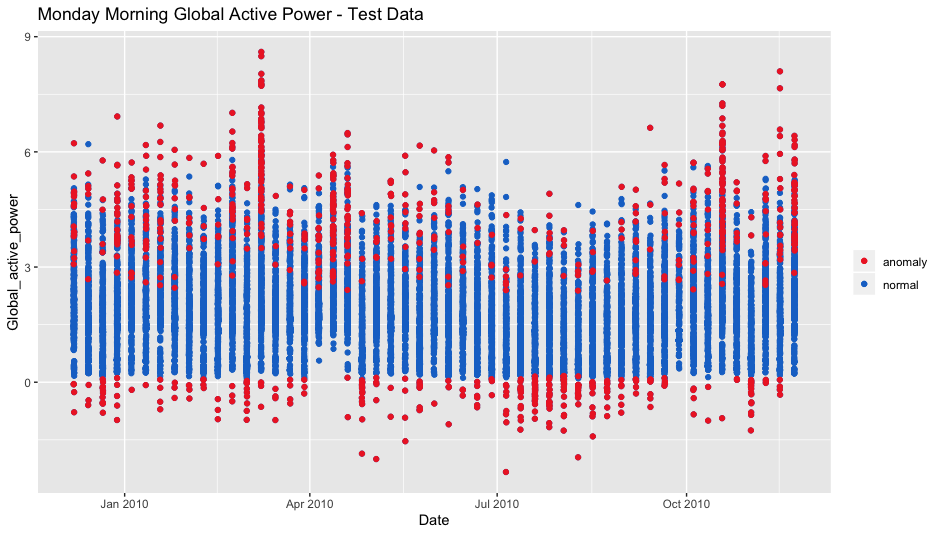
## Anomaly Detection

**Out of Range**



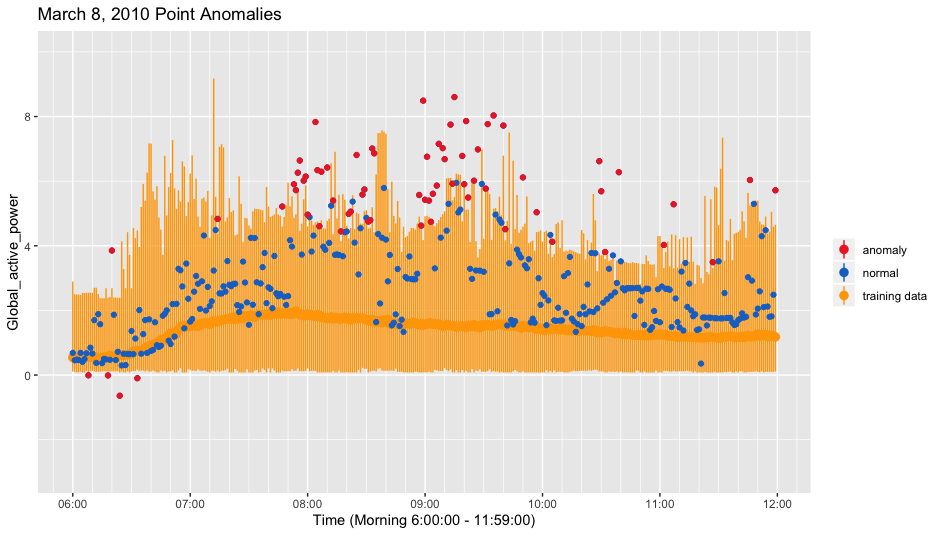
**Figure 1.** Training Data: Average global active power at each minute on Monday mornings in the training data. Error bars show maximum and minimum values for each minute.

A specific time window of Monday mornings, 6:00 am to 11:59 pm, was set over the training data to determine the average global active power for each minute. A maximum-minimum range was also established for each minute to detect all values in the test dataset which were outliers above or below the max-min range (Figure 1).

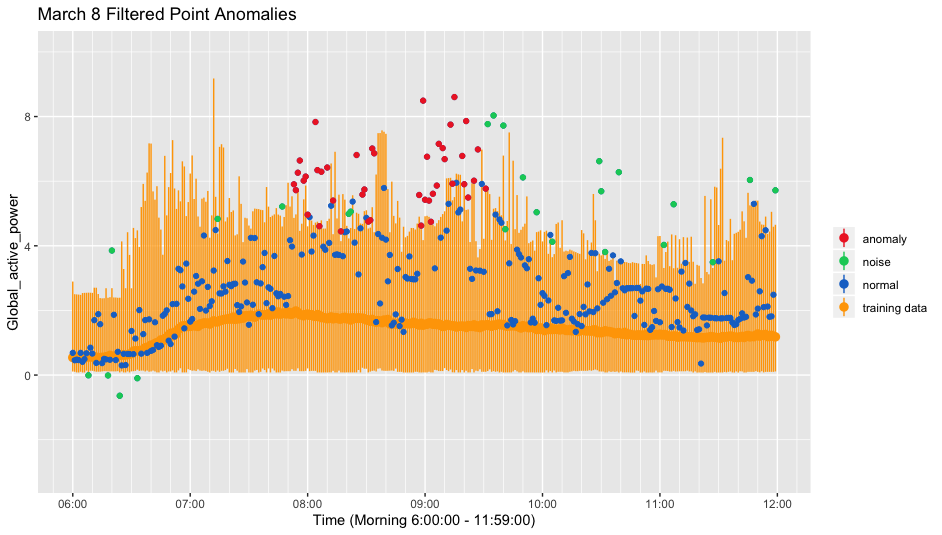


**Figure 2**. Test 1: Global active power for every Monday morning in the test 1 data set. Red points are outside of the max/min boundaries (Figure 1) defined by the training data and are marked as point anomalies. Blue points are within the normal range for global active power at the specified time.

The expected range for global active power was applied over each data set and plotted to visualize the number of outliers. Figure 2 shows every Monday morning from test1.txt and marks each outlier as a red coloured anomaly. Upon further analysis, March 8, 2010 appears to show a significant number of anomalies between ~8:00 am and ~9:30 pm (Figure 3). There also appeared to be a few outliers scattered throughout different times. An additional filter was applied to the data to distinguish noise from anomalous behaviour. If more than 50% of the observations in a 10-minute window were outliers they were marked as anomalies. All other outliers were marked as noise and determined to be insignificant (Figure 4). This technique was applied to all Monday mornings to detect anomalous behaviour from out-of-range data. Plots with anomalies for test data sets 2 to 5 can be seen in Appendix A.

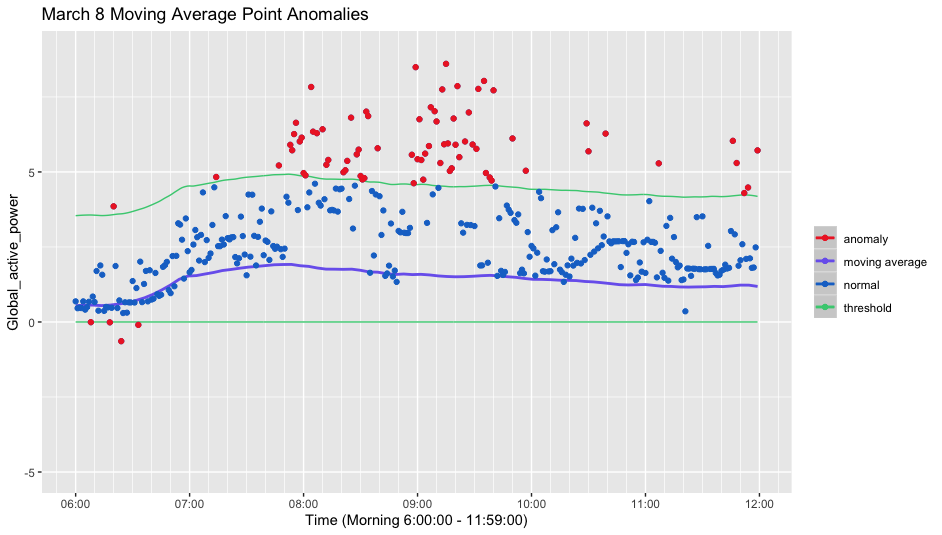


**Figure 3**. A magnified scale of Figure 2 showing global active power point anomalies for March 8, 2010, with respect to the max/min boundaries defined by the training data.



**Figure 4**. Point anomalies on March 8, 2010 were filtered for noise based on density. Anomalies were defined as more than 5 outliers in a window of ten minutes. All other instances of outliers were categorized as noise.

**Moving Average**



**Figure 5.** Test 1: Point anomalies on March 8, 2010 identified by a moving average of ~18 observations and threshold.

A moving average with a threshold was used as another anomaly detection approach. A fixed window size of ~18 observations was used to smooth the average global active power curve of the training data. A lower bound threshold was set at zero since all global active power measurements from the training data were positive. The upper bound threshold was set to +4 kW above the curve in order to capture most of the normal behavior. For comparison, March 8, 2010 was analyzed again (Figure 5) and all data points above or below the threshold were marked as anomalies. The same filter was applied as previously, where for all data points in a ten minute window, if >5 were outliers they were marked as anomalous behaviour. This technique was applied to all Monday mornings to detect anomalous behaviour from the moving average threshold.

**Summary**

Both approaches were applied to Monday morning global active power for the five test data sets, and the results are summarized in Appendix A Figures 2.1-2.5. The figures summarize every interval where anomalous behavior was identified using the corresponding approach. The out of range approach for test 1 found six instances of anomalous behavior in the dataset while the moving average approach identified nine instances (Appendix A, Figures 2.1a, 2.1b). For test 2, the out of range approach identified eight while the moving average approach identified seven (Appendix A, Figures 2.2a, 2.2b). Test 4 and test 5 performed poorly and detected a large number of anomalies, many of which are likely false positives (Appendix A, Figures 2.4ab, 2.5ab). Upon closer inspection, this could be because the average global active power for the test data has a much greater dispersion and higher average mean, as reported in Phase 1. Data Exploration. Most of the values are beyond the training data maximums so any approach utilizing an upper and lower bound are not effective. Increasing the maximum threshold could possibly yield better results, but point anomaly is likely not an effective approach for test data 4 and 5.

# Phase 2. Single Feature HMMs

In previous sections of this paper we have attempted to identify specific anomalies within our data-sets. However, in this section we will attempt to take a more binary approach and just see if we can identify whether or not our data is normal. Does it contain anomalies or not? To do this we will use Hidden Markov Models (HMMs) that have been trained to spot the difference.

Markov Models are a method used to model sequential data. They become Hidden Markov Models when underlying data used to create the Markov Model is obstructed or hidden from our view. To elaborate we are not able to view the true states of the model, we have to base our understanding of it on observational information/data. There is some relation between the true states of our model but we are unable to see those. Instead we need to utilize the produced observational data to create an educated guess about the true states of the Markov Model. In our project the observation of electricity usage is split into a number of different features but we do not know the states or how they transition from one value to the next. We can see the electricity output values but we do not have a good sense of what specifically is causing the shifts in power data. Our datasets consist of continuous electrical values each minute over the course of years. This means that we will use a continuous HMM as opposed to a discrete HMM. When constructing HMMs we compare the BIC and LogLik values to determine if our model is correct. We look for the lowest positive BIC value and the largest negative LogLik value in order to prevent overfitting.

### Log-Likelihood

Likelihood is the chance that a given model is true based on observed data. The likelihood values are from 0-1, 1 meaning that there is 100% chance that the model is true. For multiple states, the likelihoods multiply which can result in an extremely small number which is hard to compare and understand. Instead, we take the logarithm of the likelihood and add up those values. We get negative log values as the log of a number between 0-1 gives negative results. The log of 1 is 0, so the best model will have a LogLik value of 0.

### Bayesian Information Criterion (BIC)

BIC is a criterion for selecting a model within a finite set of models. The BIC is formally defined as:

Where L is the maximized value of the likelihood function of model M ie

is the observed data

is the number of data points in

is the number of parameters estimated by the model

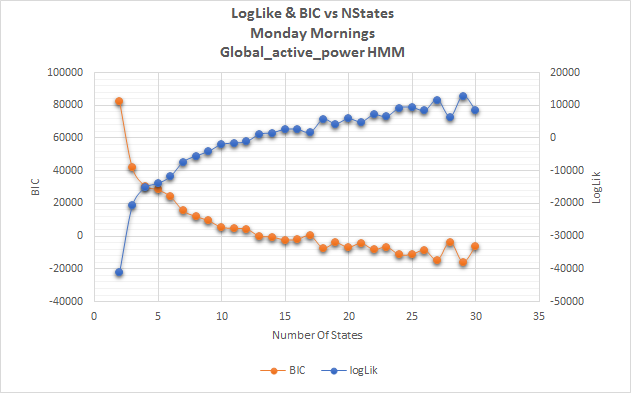
When determining the model with the lowest likelihood, we often overfit out training data. This is counteracted by looking at the BIC value which penalizes a model for having a large number of parameters in the model.

### Time Period Selection

We opted to examine Mondays and base our time window on just that day. We broke Mondays down into three separate chunks to make the data more manageable as we worked with it. The three chunks were Monday mornings, (6-11:59), Monday afternoons, (12-17:59), and Monday evening (18-23:59). We then created a single feature HMM for each of the three-time windows. Additionally, we attempted to create three additional models using three features at the same time. We met with mixed results in our attempts to encompass these extra features.

To train our three-single feature HMMs we used R and the R package depmixS4 to create our models. The single variable we picked was Global\_active\_power. We reasoned that this variable best represented the total power output of a household and it would therefore give us the clearest indication of what was happening for each household. As for our three variable HMM we used Global\_intensity as our dominant variable and the two sub variables were Global\_active\_power and Global\_reactive\_power. All three variables shared a high correlation between each other but Global\_active\_power and Global\_reactive\_power had a low correlation. This combination of relationships helped ensure that we did not over fit our model.

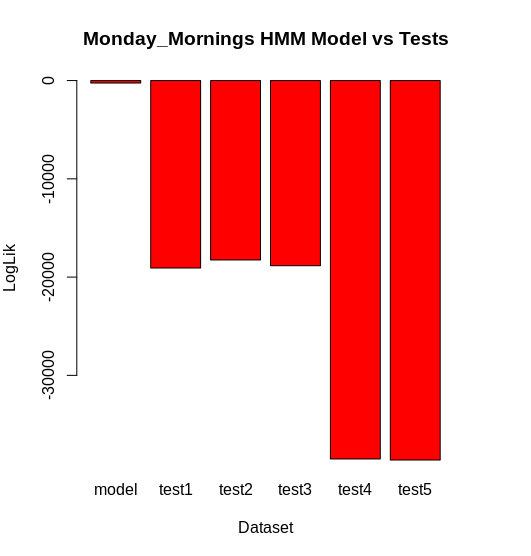
Using our training data-set we began to train our HMMs. This was done by adjusting the number of states in our model. Our goal was to find the model with the log likelihood with a negative value that was closest to zero. For example on Monday mornings we can see the result of testing two to twenty one different states.



**Figure 6.** HMM construction: A comparison of the log likelihood and BIC values of each model from 2 to 30. The best fit was at 12 nstates.

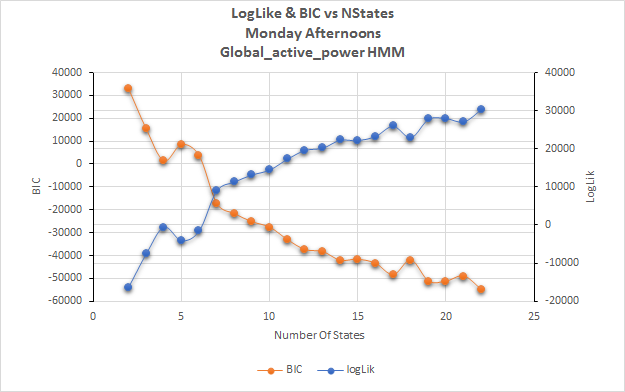
We can see that the BIC and logLik values are inversely related. Additionally, in this particular example the best model appears to be twelve states. With twelve states we have a logLik of -419.9810 and a BIC of 2665.19341. This tells us that when we begin to compare our model against the test data for this particular time window of Monday mornings, we will use twelve states in our model.

When we compare our model against the test data-sets we will expect similar logLik values as we detected in our training model if there are no anomalies. However, if the logLik values are off by a large margin, we can assume that the test data-set contains anomalies. The greater the difference the more anomalies we can expect to find.



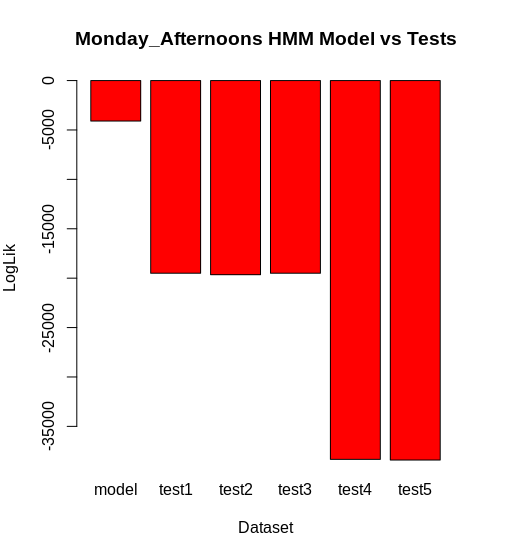
**Figure 6B.** HMM Test Comparison: A comparison of the difference in logLik between the HMM model and the 5 different test datasets.

Our model produced a logLik of -246 while tests one, two, and three had a logLik of around 19,000. Tests four and five were even further off with a logLik of approximately 38,000. These findings would indicate that there are anomalies in our test data-sets and that tests four and five in particular are comparatively worse.



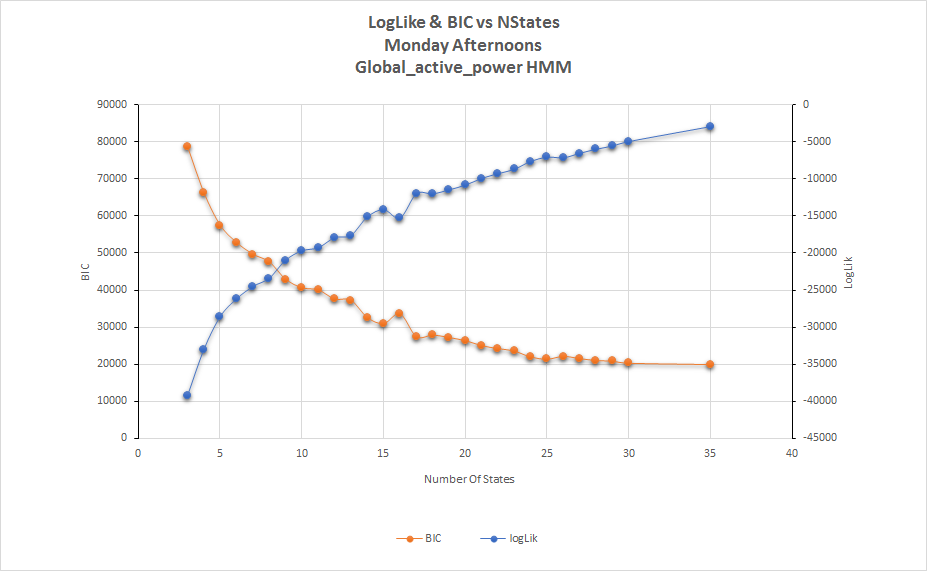
**Figure 7.** HMM construction: A comparison of the log likelihood and BIC values of each model from 2 to 22. The best fit was at 4 nstates.

Monday afternoons proved to be a very consistent time slot for Global\_active\_power. Contextually we can assume this is because power usage on a monday afternoon is fairly static with most people at work and those who are at home holding similar routines. As such our optimal model occurred at five nstates.



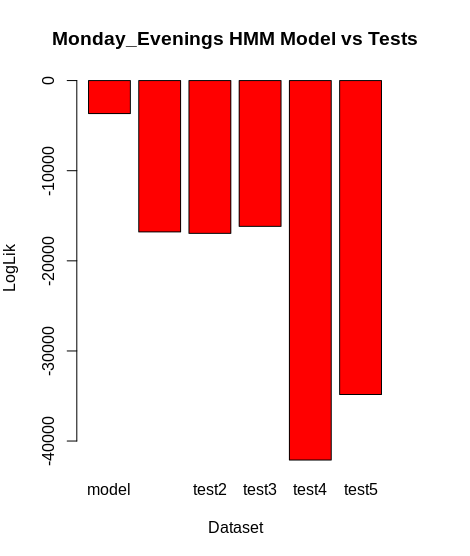
**Figure 7B.** HMM Test Comparison: A comparison of the difference in logLik between the HMM model and the 5 different test datasets.

When testing our afternoon model against the five test datasets we found a similar number of anomalies in each of the datasets as in the mornings. There are some anomalies in tests one through three and tests four and five both contained about double the anomalies of the previous three tests.



**Figure 8.** HMM construction: A comparison of the log likelihood and BIC values of each model from 3 to 35. The optimal number of states exceeded 35 but we stopped calculating there.

Interestingly where Monday afternoons were a very consistent time period without much variation Monday nights were very volatile. When attempting to optimize the model we ran out of computing power before arriving at the best number of states. As (Figure 8) indicates at 35 nstates we were trending towards a log likelihood of zero however we were likely a few states away. Computing additional nstates at that size was taking upwards of 45 minutes per state so we opted to stop at 35 and accept that our model would not be as accurate as it could possibly be.



**Figure 8B.** HMM Test Comparison: A comparison of the difference in logLik between the HMM model and the 5 different test datasets.

As we can see, (Figure 8B) it is clear that 35 nstates was sufficient to identify anomalies in our test datasets despite it not being totally optimized. Once again, the trend from the other time windows continues into the evening as there are anomalies in all five of the test sets. Although we can notice a spike in the test four dataset which is a notable change from previous windows.

### Cross Validation

After finding our optimal model parameters, we attempted to conduct cross validation. In theory, we would have conducted cross validation during our parameter optimization step but this is only if we had a large enough training dataset. Since we only have 3 years of training, we only have 3 permutations of the training set:

Table 1: Monday Morning Cross Validation Training

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Training Years** | **LogLik** | **BIC** | **Validation Years** | **LogLik** | **BIC** |
| 2007, 2008 | 223.103 | 1313.9301 | 2009 | 938.5483 | -247.6267 |
| 2007, 2009 | -667.695 | 3089.0946 | 2008 | 216.7383 | 1209.3605 |
| 2008, 2009 | -1775.386 | 5302.816 | 2007 | -103.4735 | 1852.9652 |
| **AVERAGE** | -739.984 | 3235.28 |  | 350.6044 | 938.233 |

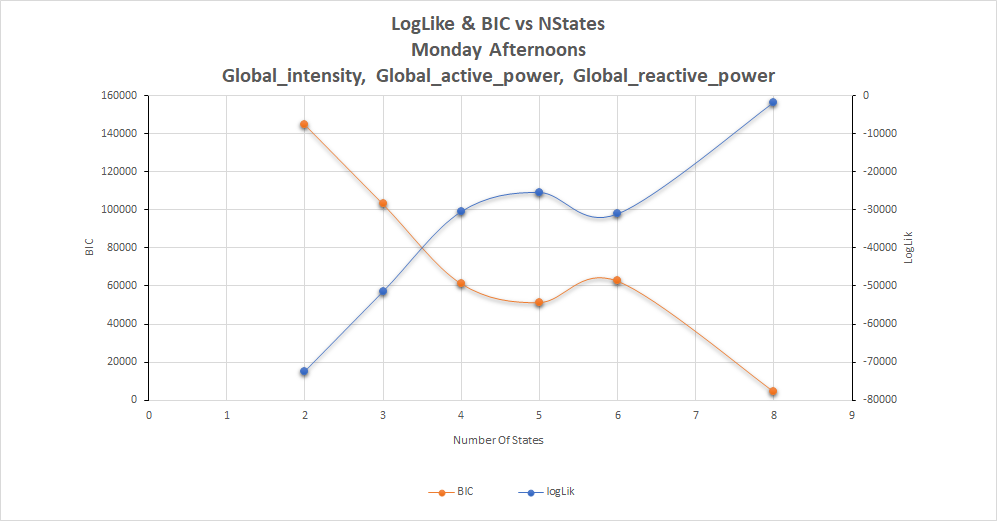
Table 2: Monday Afternoon Cross Validation Training

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Training Years | LogLik | BIC | Validation Years | LogLik | BIC |
| 2007, 2008 | -3845.847 | 8050.056 | 2009 | 2218.5168 | -4105.285 |
| 2007, 2009 | 2705.363 | -5053.685 | 2008 | 1476.5811 | -2618.692 |
| 2008, 2009 | 883.0402 | -1408.377 | 2007 | 1084.9764 | -1834.835 |
| **AVERAGE** | -85.8144 | 528.994 |  | 1593.3581 | -2852.937 |

Based on the results of the cross validation, we can see that this does not work with such a small dataset as the training dataset even fails resulting in positive LogLik value when compared to the training dataset with the total 3 years. If you take the average of the 3 training permutations, the values look reasonable but when you look at the validation results, they are completely incorrect.

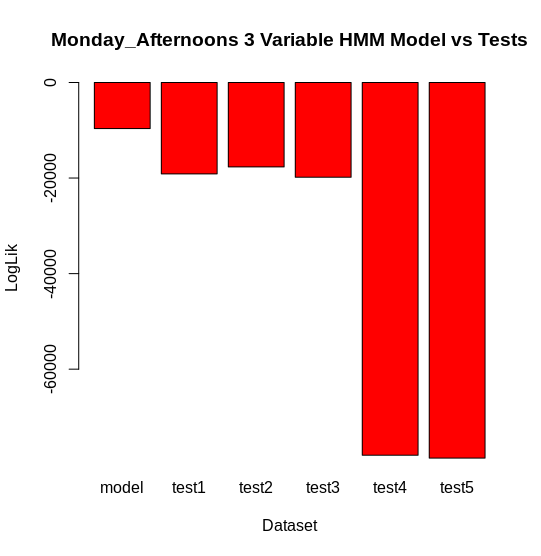
# Three Feature HMM Testing

As previously mentioned, we attempted to create a HMM using three features we used Global\_intensity as our dominant variable and the two sub variables were Global\_active\_power and Global\_reactive\_power. All three variables shared a high correlation between each other but Global\_active\_power and Global\_reactive\_power had a low correlation. This combination of relationships helped ensure that we did not over fit our model. DepmixS4 proved to be particularly finicky an it would regularly crash on certain numbers of states on one attempt while it would complete the model of the same number of states on another attempt. As such it is hard to draw any meaningful conclusions from the three feature models. That is unfortunate because the added richness of two additional features to our model would in theory make them more accurate and better suited to detect anomalies. We did have some success in crafting and testing in the Monday afternoon window.



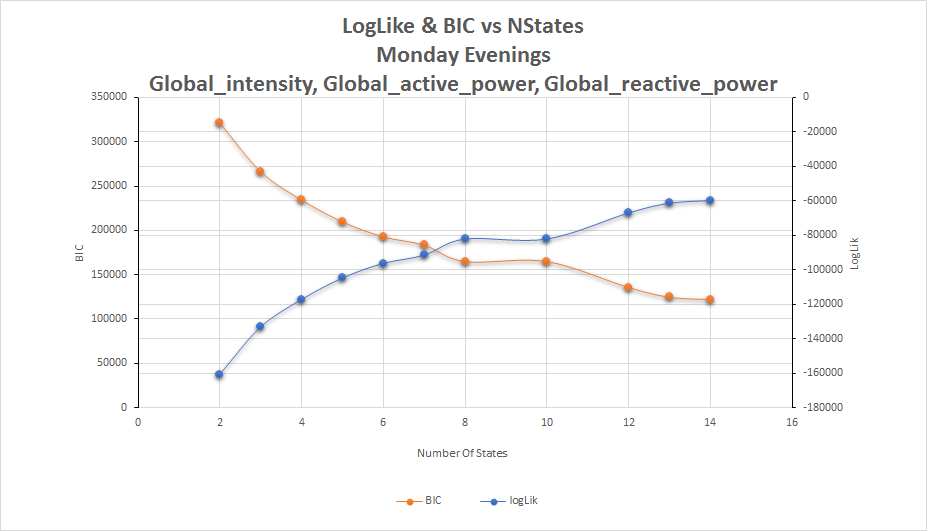
**Figure 9.** HMM construction: A comparison of the log likelihood and BIC values of each model from 2 to 8. The optimal number of states was surprisingly 6.

The results generated in crafting this model were interesting. We can see (Figure 9) that no test for seven nstates succeeded. The program crashed on seven states. Also, we can see that eight nstates produces a near perfect model however this value was unreliable and caused crashes more often than not. When testing the model against test data using the next logical choice of five nstates we also encountered strange behavior which led to a large number of crashes. So we finally settled on six nstates as this number proved to be the most stable although not the most optimized option.



**Figure 9B.** HMM Test Comparison: A comparison of the difference in logLik between the HMM model and the 5 different test datasets.

The results from our three feature HMM proved to be fairly similar to our single feature HMM for Monday afternoons. The key differences are that we are detecting more anomalies in tests four and five with this particular model and we notice a slight drop in the number of anomalies in test two.



**Figure 10.** HMM construction: A comparison of the log likelihood and BIC values of each model from 2 to 14. The optimal number of states was surprisingly 6.

For both Monday mornings and Monday evenings we encountered a similar problem (Figure 10) in that we were unable to find an optimal number of nstates before the program would crash. We attempted to test our models with these less than ideal log likelihood values and in the case of Monday mornings we did produce results (Figure 10B, 10C in Appendix) however these results were different enough from our previous models that we were suspicious of their quality. Given that tests one to three had lower log likelihood values then our training model. Given the number of anomalies we found in both our previous HMMs and our other anomaly detection methods we concluded that these results were unreliable and we have chosen to include them in the appendix.

## HMM Summary

The HMMs provide a useful tool for identifying the presence of anomalies in our datasets. The HMM is essentially the smoke detector of our anomaly detection system. They inform us of the presence of anomalies but they do not tell us where the fire is. For that we need to turn to other anomaly detection systems. The single feature HMMs appeared to be enough to detect the presence of the anomalies and while our three feature HMMs did not work perfectly they did work successfully on our simplest time window of Monday afternoons. Perhaps if we were to switch to a more robust modeling program and stop using depmixS4 we would be better able to develop our three feature HMM and gain a more accurate system.

# 

# Phase 3. Anomaly Detection Approach

## Recall

Recall (also known as sensitivity) is the fraction of relevant instances that have been retrieved over total relevant instances. Recall is a measure of ***completeness or quantity.***

## Precision

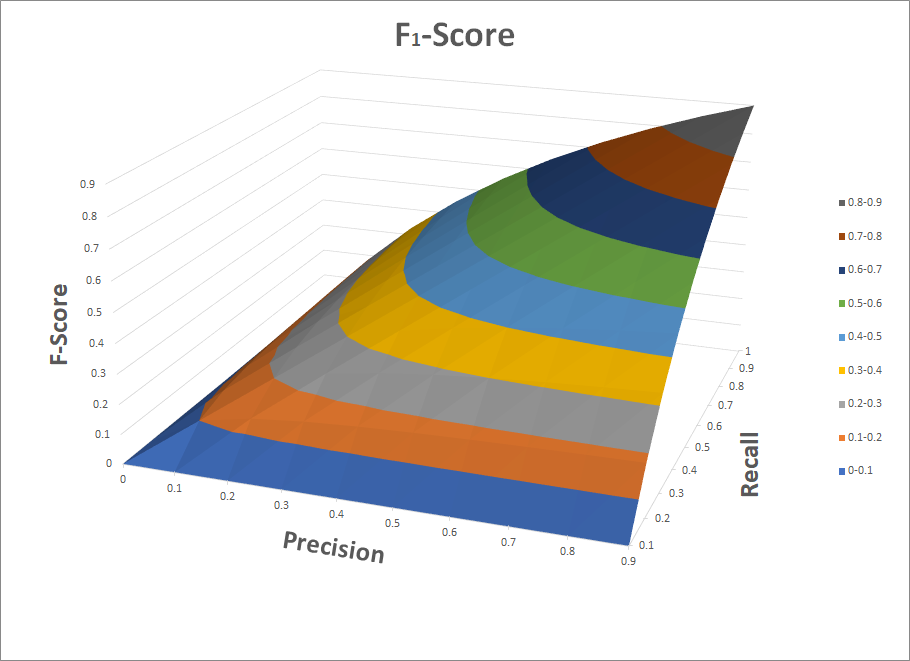
In pattern recognition, information retrieval and binary classification, precision is the fraction of **relevant instances** among the **retrieved instances**. Precision can be seen as a measure of ***exactness or quality.***

## 

## F-measure

A measure that combines precision and recall is the ***harmonic mean*** of precision and recall,

known as the traditional F-measure (or balanced F-score):



## F-Score

The F-Score is a combined measure of precision and recall.

## F2-score

F2-score weighs recall higher than precision

## F0.5-score:

F0.5-score weighs precision higher than recall

# Questions

Question: Explain what it means to optimize the balance between recall and precision.

Answer:

Scenario 1: High recall ~ 1

If you cast a large enough net out, you will be able to catch all the fish, but at the same time catch things that were not desired. If your model is very sensitive to anomalies, it may catch all the anomalies but at the same time, it will return a higher percentage of false positives thus decreasing the precision.

Scenario 2: High Precision ~ 1

If you use a single fishing rod with a very specific bait that only the desired fish will eat, you will be able to catch the correct fish but you won’t be able to catch all the fish because you are too slow. This scenario may allow you to precisely detect the correct anomalies but will miss out on anomalies due to limited resources.

Precision and recall can be viewed as two entities on a scale, increasing one will decrease the other so it is best to obtain a balance between both entities. The F-measure can be used to determine the balance between recall and precision. Obtaining the highest F-measure can be interpreted as having a balanced model. The F-measure is defined as:

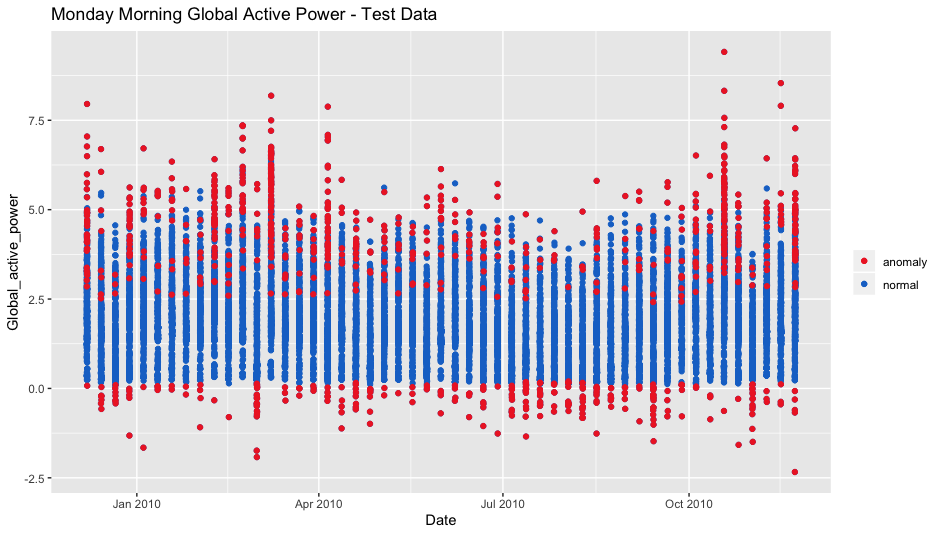
This measure is also known as the balanced F-score or F1-score, because recall and precision are evenly weighted. It is approximately the average of the two when they are close, and is more generally the harmonic mean, which, for the case of two numbers, coincides with the square of the geometric mean divided by the arithmetic mean.

The F1 measure is not the only measure that can be used. There is the F2 measure and the F0.5 measure. Which F-score is best to use is dependant on the application it is used in. Some applications need higher recall such as high security applications. Some applications need higher precision such as banning of hackers in a video game. You want to be sure that you ban the correct hackers and not regular users.

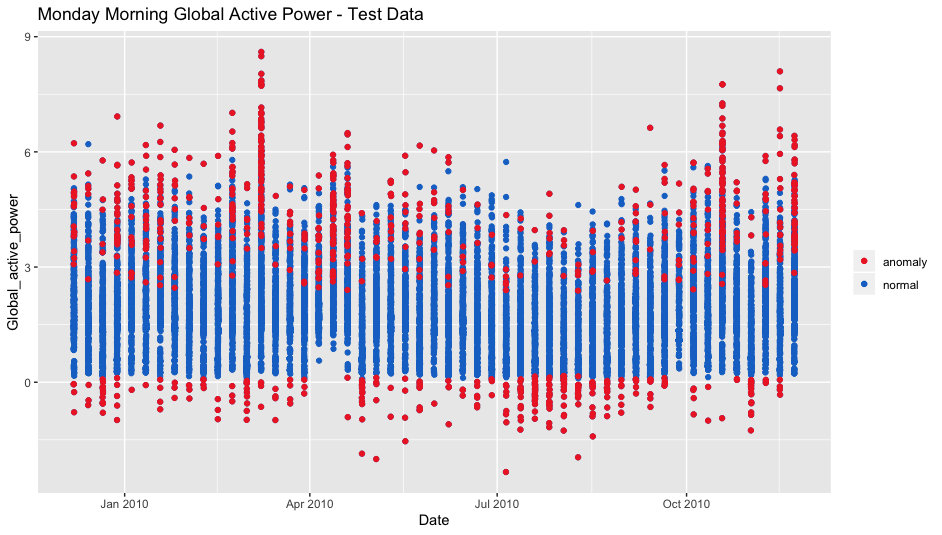
References

1. Glasser, Uwe. “318 Cyber Security Lectures.” Simon Fraser University, https://vault.sfu.ca/index.php/s/A5HO9zPeOPapwGk
2. Serrano, L. A friendly introduction to Bayes Theorem and Hidden Markov Models. Retrieved November 15, 2019, from https://www.youtube.com/watch?v=kqSzLo9fenk.
3. Seyed, Amir. “318 Cyber Security R Tutorials.” Simon Fraser University, <https://vault.sfu.ca/index.php/s/cKY6733upz3KdJ2>
4. Visser I, Speekenbrink M, “depmixS4: An R Package for Hidden Markov Models. “Journal of Statistical Software, 1-21. https://cran.opencpu.org/web/packages/depmixS4/vignettes/depmixS4.pdf

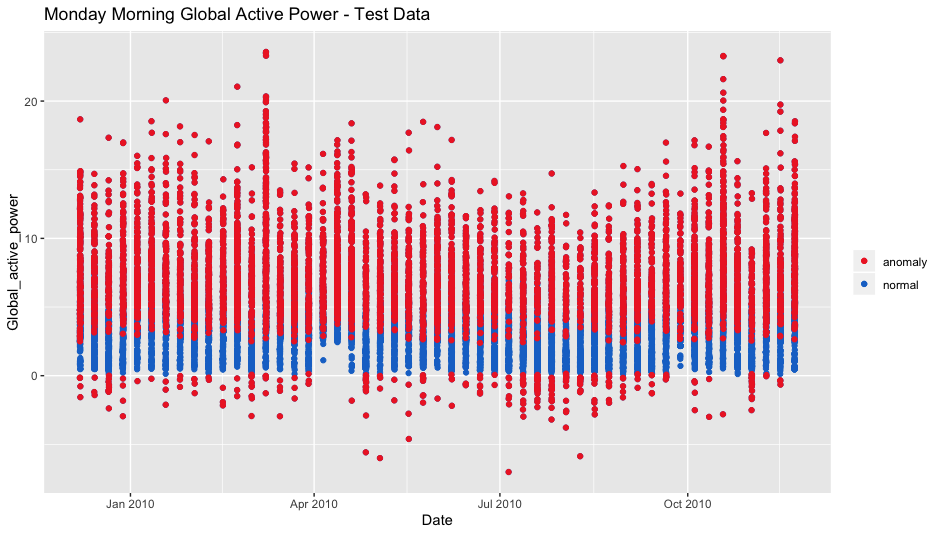
# Appendix A - Charts and Tables



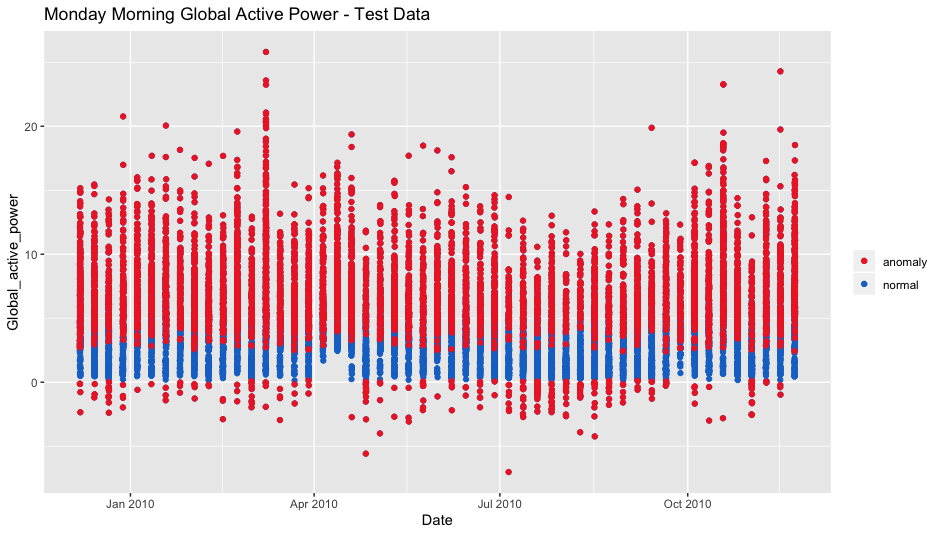
**Figure 1.2.** Test 2: Global active power for every Monday morning in the test 2 data set. Red points are outside of the max/min boundaries (Figure 1) defined by the training data and are marked as point anomalies. Blue points are within the normal range for global active power at the specified time.



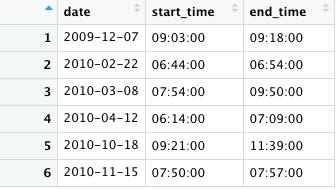
**Figure 1.3.** Test 3: Global active power for every Monday morning in the test 3 data set. Red points are outside of the max/min boundaries (Figure 1) defined by the training data and are marked as point anomalies. Blue points are within the normal range for global active power at the specified time.



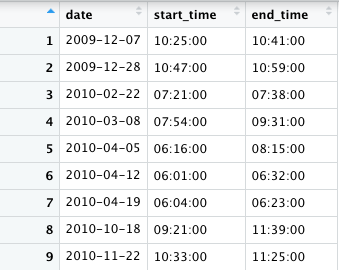
**Figure 1.4.** Test 4: Global active power for every Monday morning in the test 2 data set. Red points are outside of the max/min boundaries (Figure 1) defined by the training data and are marked as point anomalies. Blue points are within the normal range for global active power at the specified time



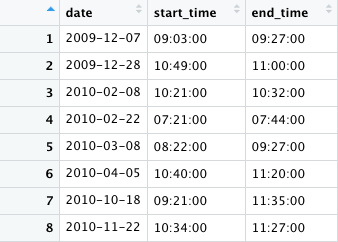
**Figure 1.4.** Test 5: Global active power for every Monday morning in the test 2 data set. Red points are outside of the max/min boundaries (Figure 1) defined by the training data and are marked as point anomalies. Blue points are within the normal range for global active power at the specified time.



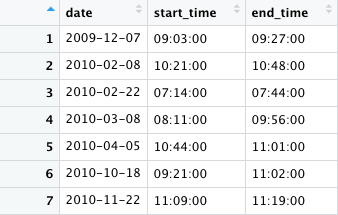
**Figure 2.1a**. Test 1: Intervals of anomalous behaviour identified as out-of-range.



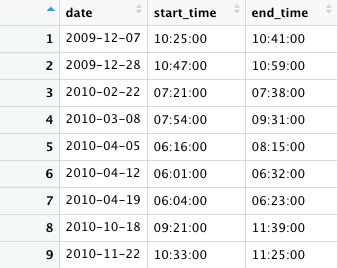
**Figure 2.1b**. Test 1: Intervals of anomalous behaviour identified by a moving average threshold.



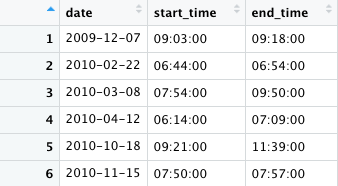
**Figure 2.2a**. Test 2: Intervals of anomalous behaviour identified as out-of-range.



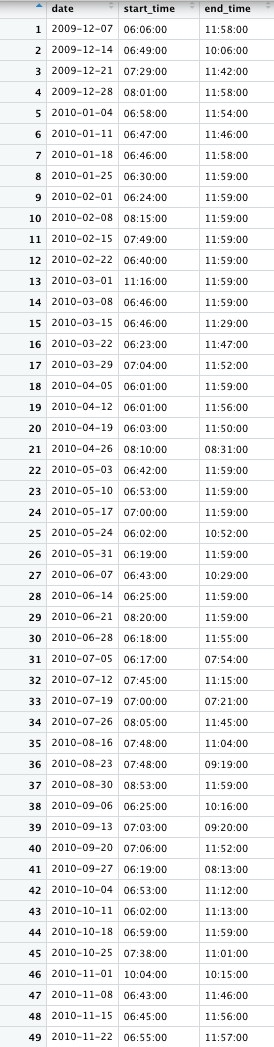
**Figure 2.2b**. Test 2: Intervals of anomalous behaviour identified by a moving average threshold.



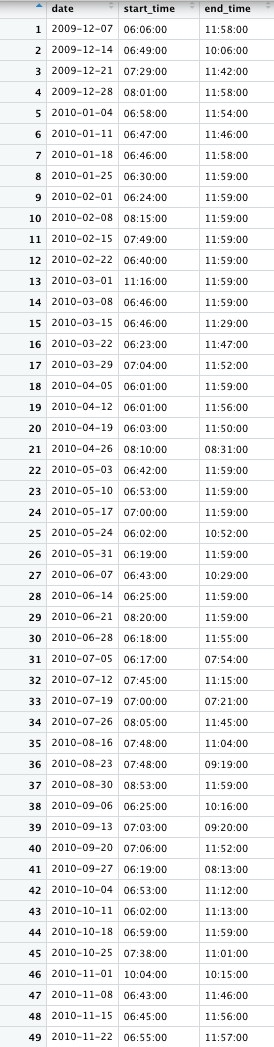
**Figure 2.3a**. Test 3: Intervals of anomalous behaviour identified as out-of-range.



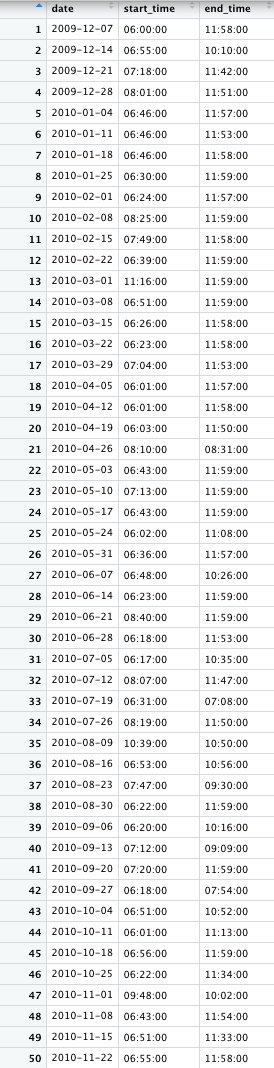
**Figure 2.3b**. Test 2: Intervals of anomalous behaviour identified by a moving average threshold.



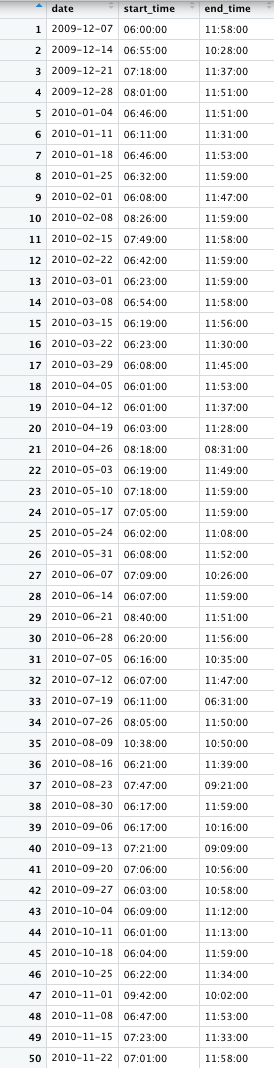
**Figure 2.4a**. Test 4: Intervals of anomalous behaviour identified as out-of-range.



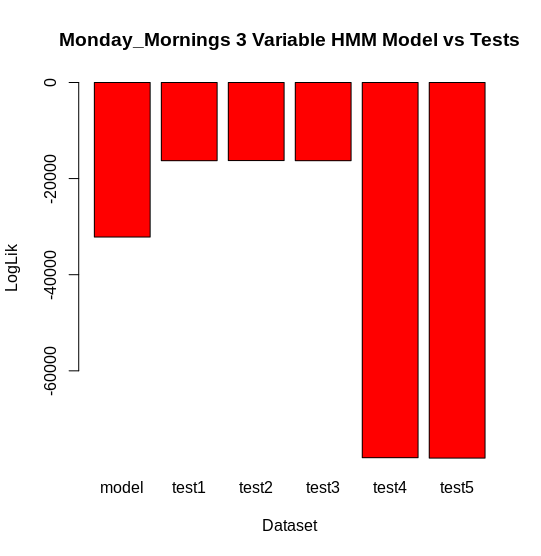
**Figure 2.4b**. Test 2: Intervals of anomalous behaviour identified by a moving average threshold.



**Figure 2.5a**. Test 4: Intervals of anomalous behaviour identified as out-of-range.



**Figure 2.5b**. Test 2: Intervals of anomalous behaviour identified by a moving average threshold.



**Figure 10B.** HMM Test Comparison: A comparison of the difference in logLik between the HMM model and the 5 different test datasets.