**Introduction**

**Problem Scope**

In this project, we are going to analyze a set of data referring to household power consumption in a specific portion of the US power grid. The aim of this project is to explore behaviour-based intrusion detection methods used for cyber situational analysis of automated control processes, such as Hidden Markov Model and anomaly detection methods. As the power grid is a critical infrastructure, not only individual consumers but also industry consumers can get affected by the fluctuation of it. We will split the whole data set in to parts by specific period of times, such as Friday nights and weekends. The challenges we face is that the data is not guaranteed to be prefect, lack of ground truth, no label added, and various type of anomalies. These defects increase the difficulty of distinguishing anomalies from noise and simple error. In the rest of this report, we are going to list our approaches and methods used in the research such as moving average method when doing anomalies identifying and Hidden Markov Model when doing data analyzing. Finally, the paper will go over the result of the approaches we used and discuss some problems occurred and encountered through the course of completing this project.

**Our Approach and Methods**

Dataset Processing

When going through the dataset we are going to use for research, there are many preparations to be done before we start our analysis. The imperfection of the dataset needs us to remove noise and null values from it. Noise data would result as a point anomaly incorrectly and lead to an inaccurate Hidden Markov Model. Null data would increase the inconvenience of further data processing.

After we clean the dataset, we decided to split the dataset into parts based on time period to make the trends more clear and easier for further analyze. The dataset contains power consumption data on each minute in 3 years. We assume there would be some behaviour-based trends on seasons, trends on specific day of week, and trends on specific time of the day. To test our assumption, we decided to split the dataset into:

There are 9 columns of data, which are:

* Date: Date as dd/mm/yy
* Time: Time in the day as 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)
* Global\_intensity: Household global minute-averaged current intensity (in ampere)
* Sub\_metering\_1:
* Sub\_metering\_2:
* Sub\_metering\_3:

There are 1 training dataset and 5 testing datasets provided. The training dataset contains 1556444 rows of data and each test dataset have 518816 rows of data.

Use *as.POSIXIt()* function to get the date format set as *%d/%m/%y* , and use *weekdays()* function to set the day of week on each day for later split.

Finding Point Anomalies

As we suppose all the data in training dataset are normal. Picking the minimum and maximum value of each feature we work on, any value above the maximum or below the minimum in test datasets would be labeled anomaly.

After we have our anomalies benchmark, we are going to use moving average to identify anomalies. Moving average is used to smoothen the curve and reduce the affect created by the noise value. We decided to use 7 observations as our window size, which means take the average of 7 nearby observations, and compare the moving average with the threshold we set by maximum and minimum. The size of window affects the smoothness of the curve and the intensity of avoiding noise. If the size too big, the intensity of avoiding noise would be high which may lead to miss some anomaly as normal, if the size too small, the intensity of avoiding noise would be low and lead to normal value identified as anomaly by mistake.

Hidden Markov Model

Hidden Markov Model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (i.e. hidden) states. (Wikipedia, 2018). In this project, we only have the observations of electricity consumption features, neither the states nor transactions are known or visible, therefore we decide to use HMM to find proper amount of states and transaction through them.

As the observations are continuous value of electricity consumption features, we use continuous HMM rather than discrete HMM (which needs the observations to be isolated and no numerical relationship to each other).

Due to the nature and ourselves experience, we decided to build our time windows by splitting the data into weekdays, weekends, Sunday mornings, Sunday nights for each season. In summer and winter, we assume people would consume more electricity as using air conditioner or electric heater. And due to normal working schedule and lifestyle of people, weekdays and weekends may have different preference of using electricity. Then, Sunday morning is considered as the middle of the weekend and Sunday night is the interact of weekend and weekdays, we are curious about if there would be any different model due to these two periods of time.

We designed our program to train our model by number of states from 2 to 15 to find the most proper HMM by comparing the log-likelihood and BIC. Models with high log-likelihood and low BIC would be regarded as the suitable solution to do further test with test datasets.