Analysis and optimisation of building management systems by forecasting energy usage and detecting outliers in real time

David Sal

3rd Year Project Interim Report

Department of Electronic &
Electrical Engineering

UCL

Supervisor: Dr Ryan Grammenos

15 December 2021

I have read and understood UCL's and the Department's statements and guidelines concerning plagiarism.

I declare that all material described in this report is my own work except where explicitly and individually indicated in the text. This includes ideas described in the text, figures and computer programs.

This report contains 15 pages (excluding this page and the appendices) and 3954 words.

Signed: _______David Sal_________ Date:____15 December 2021_____

Analysis and optimisation of building management systems by forecasting energy usage and detecting outliers in real time

David Sal

Energy consumption forecasting for buildings plays a significant role in their energy efficiency. Accurate forecasts and proper energy management strategies allow for decreased energy usage that reduces the environmental effect of the building and presents economic benefits due to decreased operating costs. Predicting energy load accurately is very difficult due to multiple exogeneous factors that affect the energy behaviour of the building such as weather conditions and occupancy schedule. This work aims to find the relationship between exogeneous factors and energy usage and develop an accurate forecasting model for energy consumption whilst being able to detect outliers in real time.

1 Introduction

Energy has always been one of the most precious resources on earth and in the past few decades it has seen an exponential increase over time. According to [1] 40% of total energy use in Europe is made up by buildings whilst emitting 36% of the total CO_2 emissions. According to [2] buildings use 20% more energy than required due to human errors, malfunctioning equipment, and faulty control systems. Using a building management system these errors can be identified and resolved to avoid wasting energy. This surge in energy demand will continue in the future due to growing population and spread of energy demanding technology, thus using energy efficiently has caught the attention of many researchers. Forecasting energy consumption can benefit energy efficiency significantly. Accurate forecasts and proper energy management strategies would allow for decreased energy usage thus reducing the environmental effect of buildings and presenting economic benefits due to decreased operating costs. With the advances of sensing technology most buildings are equipped with sensors that measure and store energy data. These sensors generate a vast amount of data which can be analysed to gain insightful information that could improve the energy efficiency of the building. Using historical energy data of buildings, a model can be created which can help us assess the dynamics of the building, forecast future consumption, and find outliers. However, the multiple exogenous factors influencing the energy load such as weather conditions and occupancy schedule, make accurate predictions very challenging. Using complex machine learning algorithms and neural nets this challenge can be overcome

1.1 Literature review

As mentioned above, energy load forecasting has been of great importance and gained lot of attention in recent years. Many forecasting techniques emerged in the past decades that can accurately predict energy load based on historical data. In this work we will focus on time series forecasting methods given our time series data. We can divide the analysis of time series data into 2 important steps: First we need to obtain the structure of the underlying pattern of the given dataset. This can be done by decomposing our time series data into three components: trend, seasonality, and residual. The trend is the general movement of the dependant variable. The seasonality is the periodic fluctuation of the variable whilst the residual accounts for the remaining unexplainable parts of the variable. The more complex part of creating a forecasting model is the second step where we try to fit a model to make predictions for the future. We will discuss the most popular forecasting methods by presenting case studies for each. A more detailed description for each method can be found in the "Background theory" section.

One of the most popular forecasting methods is using Artificial Neural Networks (ANN). Nizami and Al-Garni give an extensive explanation of ANN in their paper [3]. In their work they presented a two-layered feedforward ANN to predict electrical energy consumption. For the predictions they used multiple exogenous factors as inputs to the NN. The results were compared to a regression model where the ANN significantly outperformed the regression model. Model adequacy is tested using visual inspection and the chi square test whilst for model validation predicted values are compared to values outside the training dataset. However, there is no further investigation of how inputs and outputs are related making the algorithm behave entirely like a black box. In a different case study Karatasou et al. [4] further improves the accuracy of ANN method by combining it with statistical methods such as hypothesis tests and cross validation as guidance for model selection. Two datasets are used for this research both consisting of hourly measurements for a period of over a year. The inputs to the NN are made up of exogenous factors, timestamp, and values from the previous timestamps to predict hourly load profiles using a feedforward neural network for modelling. This paper puts lots of effort into validation of the model and gives great attention to input selection by finding the relationship between inputs and output. However, data pre processing was neglected making the model prone to fail on datasets with significant outliers.

Another forecasting method is the Autoregressive Integrated Moving Average (ARIMA) technique. Newsham and Birt [5] proposed an ARIMAX (ARIMA with eXogenous input) model for predicting power demand for an office building. As exogenous input only the building occupancy was used. The dataset used consisted of hourly measurements over a period of 79 days with multiple missing values. These were imputed by the mean of values for that hour from that week. Whilst the ARIMA model presented gives accurate forecasts this work only focuses on how exogenous input can increase the accuracy the model whilst not exploring any other processes in further depth.

Support Vector Machines (SVM) are a further model frequently used for energy load forecasting. Xuemei et al. [6] proposed a Least Square SVM to forecast cooling load of an office building using exogenous factors. Whilst the results are superior in comparison to a back propagation ANN this work does not investigate the relationship between input and output values. A similar case study presented by Fu et al. [7] SVM was used to forecast energy demand for different subsystems. This work compares SVM to ARIMAX, ANN and a further method. Whilst the paper shows the superiority of SVM related to the other methods it lacks the same analysis as in [6].

Lastly, an emerging forecasting method is using hybrid models where 2 or more methods are combined for more accurate predictions. The combined methods can complement the advantages of the individual techniques creating a superior forecasting method. Zhuang et al. [8] combines ARIMA with ANN, Nie et al. [9] combines ARIMA and SVM in their work. Both cases show that the hybrid method outperforms both initial methods whilst many researchers achieve similar results using different combination of forecasting methods.

Overall, there are multiple techniques that can be used for series predictions which all result in accurate predictions. Depending on the case study and the dataset specific methods might outperform the others thus multiple forecasting methods need to be examined before making a choice. Despite the studies in the literature review producing acceptable results multiple weaknesses can be identified. The most common one is not assessing their model using interpretability techniques for making the black box process more explainable that would lead to a more robust model and could obtain hidden information in the dataset. Further weaknesses were not cleaning the dataset from inconsistent values and using weak imputation methods for missing values.

1.2 Objectives of this work

- Reproduce reusable results from the MSc Project study conducted by Polyxeni Kalliga on the same dataset for validation purposes.
- Create an energy forecasting model that can accurately predict values for a specific timestamp based on previous datapoints and influencing factors such as outdoor temperature, humidity, day of the week etc previous datapoints. For this, multiple forecasting methods need to be explored such as the ones mentioned above in the literature review section.
- Finding the relationship between energy load and exogenous factors by firstly implementing a prediction model and then explain the relationship between the predictors and the response variable using interpretability techniques. For this interpretability methods such as global/local surrogate need to be explored.
- Implementing an outlier detection method that compares the predicted value to the actual value and decides if it should be considered as an outlier. For this we will explore multiple outlier detection methods such as sliding window and other anomaly detection methods on streaming data from [10].

2 Background theory, methodology and results

In this section we introduce the relevant background theory to this report and show the work conducted and the results up to this date.

2.1 Background theory

This section gives an overview of the theories required for this project.

2.1.1 About the dataset

This work analyses building data of 2 office building blocks located on the Hursley IBM industrial site. The data used for this study is half hourly energy demand over a period of two years from January 2018 to December 2019. The building data consists of electricity, heating, cooling, and water capacity measurements for the Hursley House block and electricity, heating and water capacity measurements for the D East block. The dataset has multiple outliers and missing values.

2.1.2 Data pre-processing

Data pre-processing is the most important step of any data analytics process. Before performing any actions on a dataset, one needs to explore it identify patterns and clean the data from any inconsistent values. The steps of data pre-processing are outlined in the following sections.

A) Exploratory Data Analysis (EDA)

The first step of data pre-processing is to explore the data by visualising points in the dataset. Depending on the nature of the data this can be done by using statistical visualizations such as box plots, scatterplots, pie charts or histograms. This step helps us to familiarize ourselves with the data, identify hidden patterns, detect outliers, and make assumptions about the dataset.

B) Identifying inconsistent datapoints

Inconsistent points are outliers and missing points in the dataset. Outliers are unrealistic values that can be caused by human errors, malfunctioning equipment, faulty sensors, or control system with incorrect configurations. They can be detected by visualisation or by probabilistic methods that determine how likely the existence of the specific datapoint is. After choosing a probabilistic method for likeliness calculation and defining a threshold for what probability shall be considered as an outlier they can be easily identified. The most common probabilistic method is calculating the Z-score which can be obtained by subtracting the mean of the dataset and dividing by the standard deviation as shown in equation (1):

$$Z = \frac{x - \bar{x}}{s}$$

The larger the absolute value of the Z-score the more likely it is that the observation is an outlier. The threshold is usually set to be 3. Given the sensitivity of the mean to outliers this method can be easily skewed thus the mean can be replaced by the median or the mean of only the numbers inside the interquartile range which brings us to the next probabilistic method.

The interquartile range (IQR) can be obtained by placing all observations in ascending order and subtracting the value at 25% of the data length from the one at 75%. The value at 25% is denoted as Q1 whilst the one at 75% is denoted as Q3.

The outlier fences can be defined by the following equation:

$$IQR = Q3 - Q1$$

$$Minimum = Q1 - IQR * c$$

$$Minimum = Q3 + IQR * c$$

Where c is a constant and usually chosen to be 1.5.

C) Cleaning the dataset

Cleaning the dataset of inconsistent points is of great significance before performing any further action on it. Understanding the reason behind an outlier is of great importance since they might contain crucial information about the building management system. Deciding what to do with an outlier can hugely affect the results of the following steps of data analysis. Removing outliers can lead to information getting lost whilst keeping an outlier can distort statistical analysis on the dataset due to most of those techniques being sensitive to outliers. Furthermore, in the case of time series data outliers cannot be simply removed as that would result in a missing point. Missing points for time series data is a further issue as we need to decide what value to impute when filling them. Common methods are to fill them using the mean of the dataset or based on adjacent values. When imputing artificial values, the challenge is to not distort the pattern of neighbouring values and not to change statistical features of the dataset either.

D) Data enrichment

Data enrichment is a process where we add additional information to our data that might be an influential input to the following machine learning (ML) models or Neural Networks (NN). For energy forecasting this information might be weather conditions (temperature, humidity, solar radiation, windspeed), occupancy, week of the day, month, season, holidays etc. Using these the models can be improved which result in more accurate predictions as proven by many works mentioned in the literature review.

E) Transforming the dataset

Before moving on to creating an energy consumption model the dataset needs to be transformed into a form that can be fed into a ML algorithm or used for training for a NN. This requires the dataset to be in one data frame with all the other inputs with normalized values.

2.1.3 Time series forecasting methods

Artificial Neural Network – ANN

ANNs are modelling techniques that are trying to mimic the human brain. Just as a biological brain it is made up of large amount of processing units namely, neurons. These neurons are connected in parallel and process information using the weights attached to each neuron. These weights can be obtained by feeding historical data into the model whilst adjusting the weights so that the predicted output gets closer to the desired output. This process is called training of the network and is done by trying to minimize the squared error between the desired and predicted output.

Autoregressive Integrated Moving Average – ARIMA

It is one of the most general forms of energy consumption forecasting methods. The idea behind it is to convert the time series dataset into a static one thus making it "integrated". This is done by a differencing process. The predicted value can be obtained from previous instances of the dependent variable from previous values of the error term between the predictions and the actual value.

The ARIMA equation is as follows:

$$\begin{aligned} Output\ (y) &= constant + weighted\ sum\ of\ previous\ instances\ of\ y \\ &\quad + weighted\ sum\ of\ previous\ values\ of\ the\ error\ term\ e \end{aligned}$$

The previous instances of y are referred to as the "autoregressive" part whilst the error terms are referred to as the "moving average".

Support Vector Machines – SVM

SVM are algorithms that analyse data for classification and regression analysis. It is mapping inputs to a high dimensional feature space and distribute the data into two sets whilst trying to find the largest margin between them. It applies the statistics of support vectors to classify data and categorize unlabelled datapoints.

2.2 Methodology and results

This section shows the current progress of the work and the obtained results until this point.

2.2.1 Exploring the data

Given the importance of EDA as described above we start off by exploring the data and visualizing it using different plots. Plotting the HH Electricity data we get the following graph:

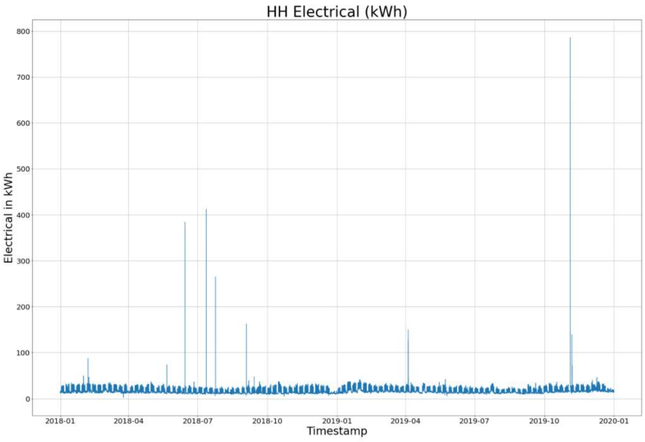


Figure 1 - A plot of HH Electrical data

As we can see in figure 1 there are clear outliers that need to be dealt with. For further analysis all points that are clearly anomalous have been visualized and analysed by searching for them in the dataset and assessing the neighbouring values. A few examples of these can be seen on Figures 2-3:

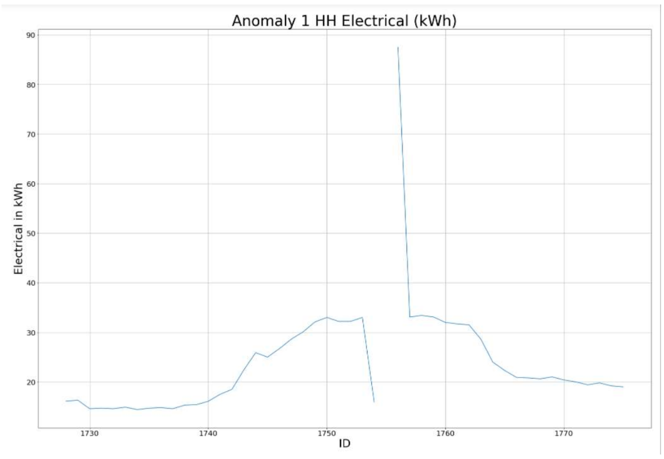


Figure 2 - A plot of an anomaly point

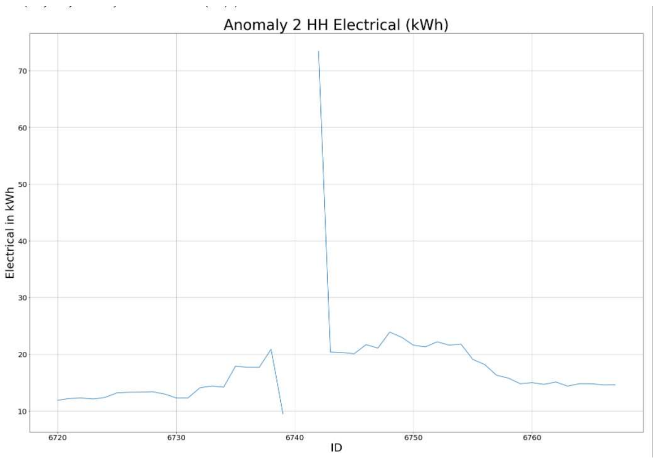


Figure 3 – Plot of an anomaly point 2

We get the suspicion of anomaly point following a specific pattern. For HH Electrical it seems as if for every outlier we have a sudden drop, a missing value which is followed by a large outlier. When further investigating this idea, we find that our suspicion seems to be correct. We also find that the energy usage over days follow a pattern too as can be seen in figures 4 – 6:

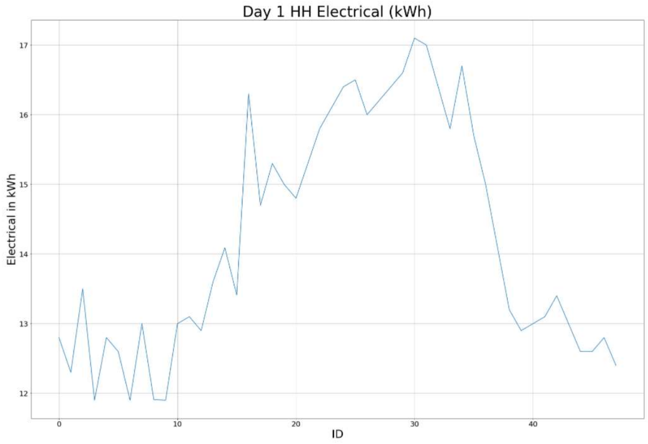


Figure 4 - Plot of HH Electricity Day 1

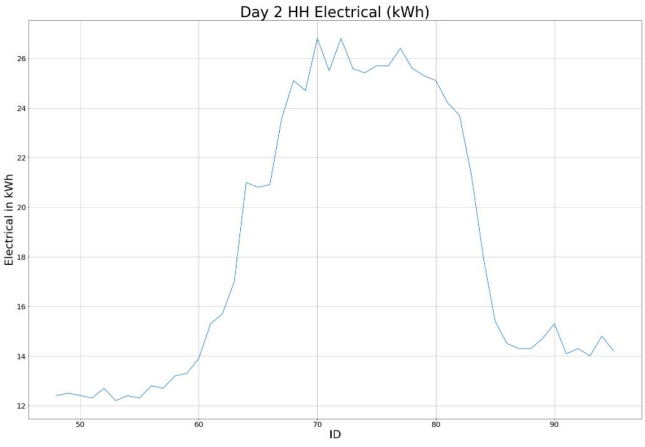


Figure 5 - Plot of HH Electricity Day 2

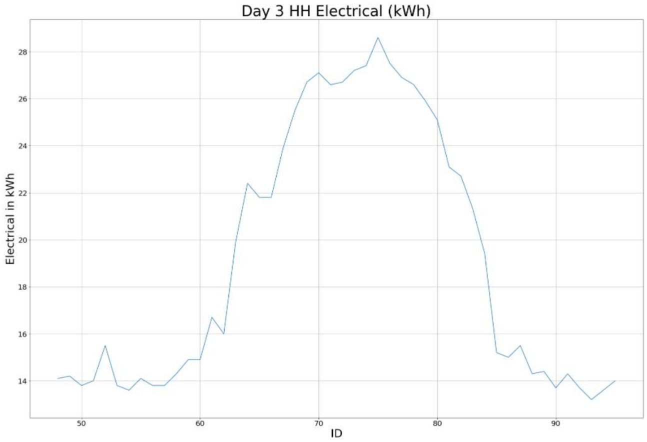


Figure 6 - Plot of HH Electricity Day 3

When assessing day 164 and 165 a very large gap was found with missing values up to half a day. Repeating this process for other features we get to similar findings.

Overall, it seems like the data is compromised by many outliers and missing values. Luckily, it seems like they all follow a similar pattern for given features with all of them having a logical reason of how they originated and propose a way to deal with them. By identifying statistical outliers using the modified Z-score method described in the background theory section and by carefully observing lots of these outliers for all features for both building blocks the following patterns can be found:

Electrical:

- Low value – missing value(s) – Outlier
- Missing values – outlier
- Low value – outlier
- Low value - high value

Heating:

- Peak – missing value(s) – sag
- Missing value(s) – High value

Cooling

- Peak – sag

Water capacity

- Missing value(s) – outlier

Where outliers are the ones that were classified as those using the modified Z-score approach, low values are such values that unexpectable drop compared to the previous value. High values are those that unexpectable increase compared to previous values. Sags and Peaks are negative and positive outliers that exist in pairs.

All these patterns suggest that the way to deal with them is to apportion the sum of affected values between the compromised points. This way the total energy is not changed and replacing outliers can be explained by findings from the EDA process

When comparing the results of the EDA with those of the previous student working on this project the findings were the same.

2.2.2 Cleaning the data

Given the information extracted from the EDA process, cleaning the data consisted of searching for the given patterns discovered in the EDA and apportioning the sum of the inconsistent values across the affected region. When doing this the results obtained were like the graphs in the report of the previous student. However, when translating the flowchart from her report into code for validation purposes the output is not as expected.

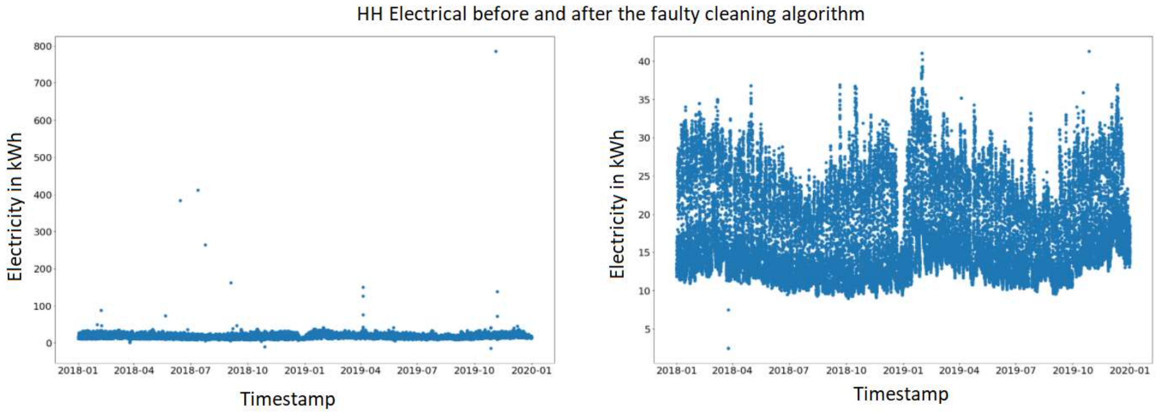


Figure 7 - Output of Flowchart code for HH Electrical

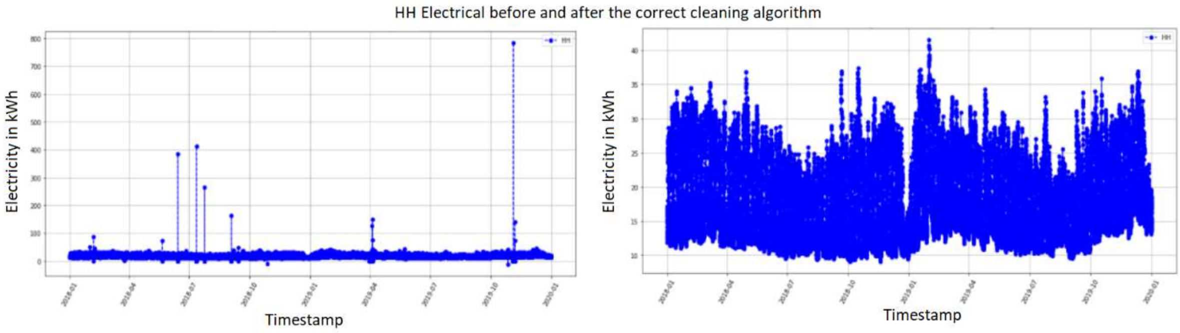


Figure 8 - output of correct cleaning algorithm

When analysing differing points between Figure 7 and 8 one can replay the flowchart algorithm by hand. When analysing the output, the results suggested that the flowchart describing the cleaning algorithm was incorrect. This was confirmed by the student when discussed in a meeting.

This issue could be solved easily by adding parts to the code that were missing from the flowchart.

The changes were the following:

- Cleaning sag peak patterns in feature 2
- Cleaning sag – zeros – peak in feature 3
- Cleaning sags in feature 2 through linear interpolation

- Cleaning feature 6 where we have zeros – outlier that should be no values – outliers but it has been manually adjusted in the original dataset.

After adding the parts described above the code properly cleaned the dataset and the results agreed with the ones presented in the previous student's paper.

2.3 Analysing the cleaned dataset and transforming it

Once the dataset is cleaned it is beneficial to analyse it again in more depth to extract hidden knowledge which might aid the structuring of the dataset for the transforming part. Since we are interested in forecasting the energy in the following sections only Electrical data will be analysed.

When plotting energy consumption averages for every hour in a day we get the following result:

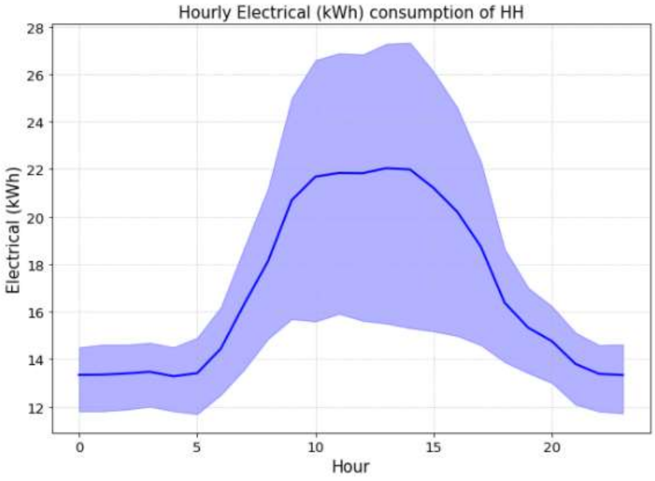


Figure 9 - Hourly trend of HH Electrical

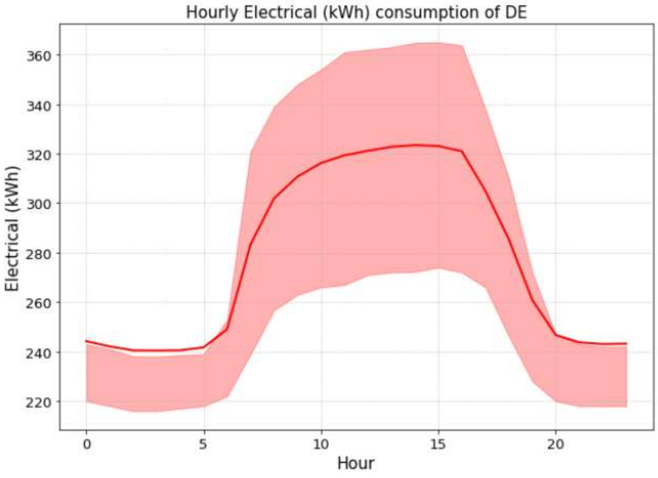


Figure 10 - Hourly trend for DE Electrical

As we can see both building blocks exhibit a significant dependence on the hour of a given day suggesting that it needs to be used as an input parameter for forecasting models.

Repeating this for day of the week we get the following plots by summing energy loads for every day and grouping the days together:

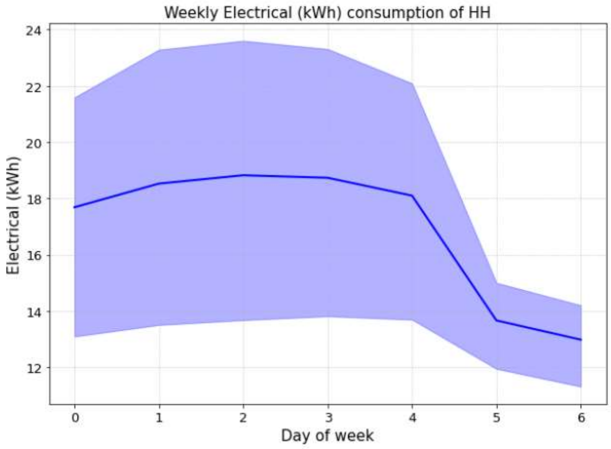


Figure 11 Weekly trend for HH Electrical

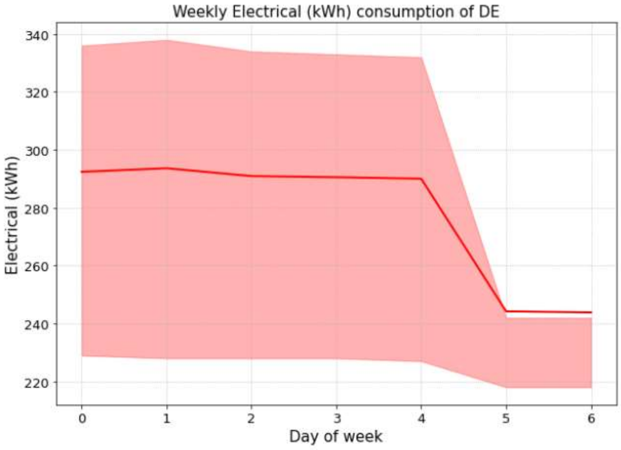


Figure 12 - weekly trend for DE Elec

As Figure 11 and 12 suggest, the day of the week is also an important predictor for energy load forecasting.

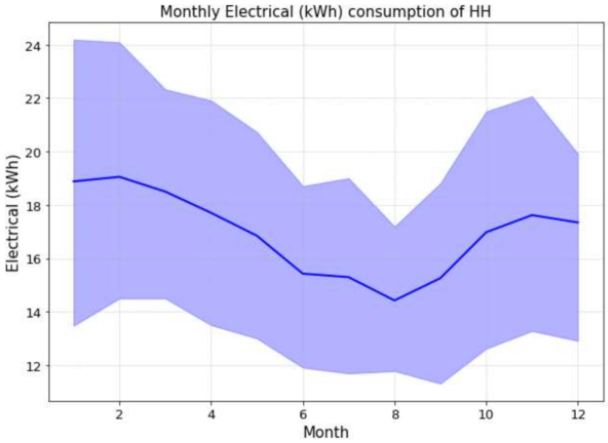


Figure 13 - monthly trend for HH Elec

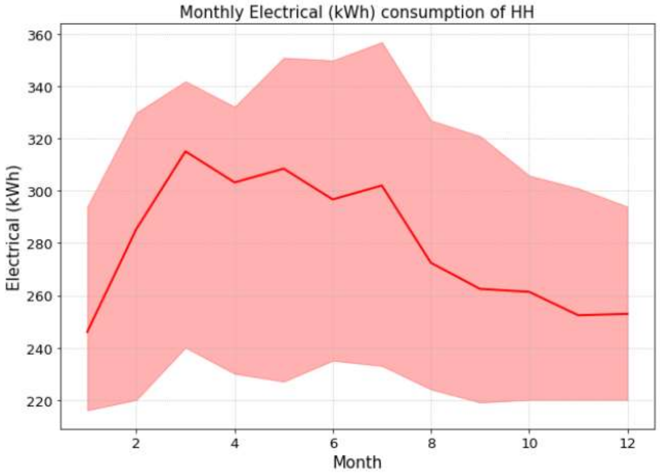


Figure 14 - Monthly trend for DE Elec

Figure 13 and 14 show us the monthly trends for HH and DE respectively. Whilst seeing obvious trends it is interesting to note that the two trends are reversed. For HH the electrical consumption drops during summer due to decreased heating energy demand whilst for DE block it increases. This can be explained by the DE building being more modern and having an HVAC system whilst cooling in the HH building block relies on natural aeration.

Overall, every plot agreed with the ones shown in the previous students work.

2.4 Data enrichment

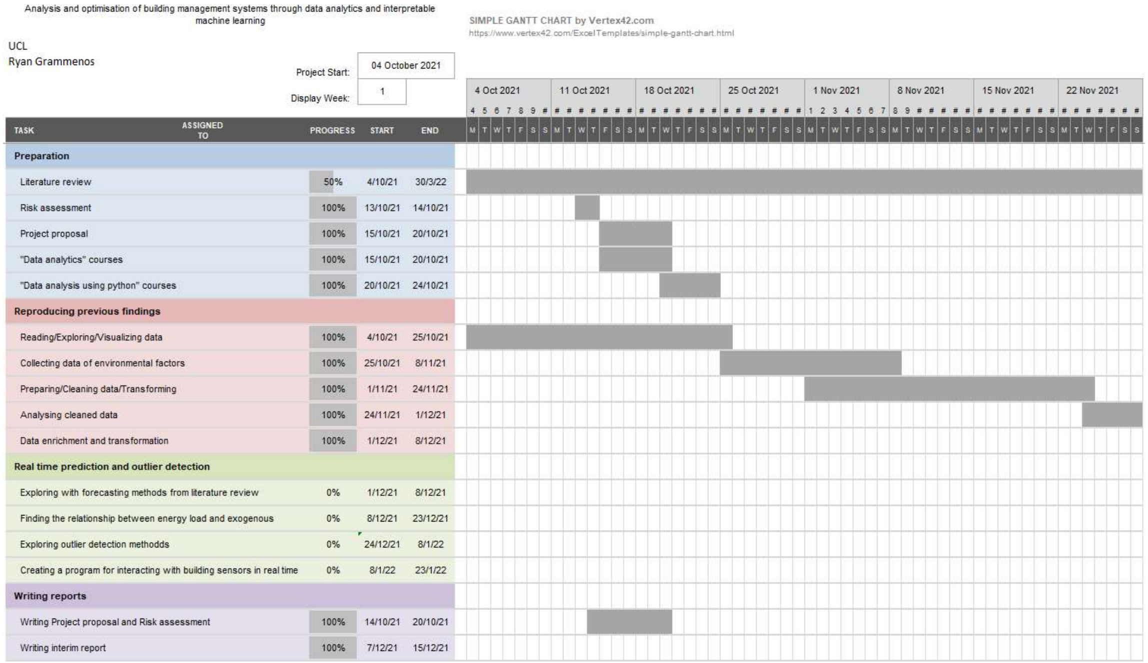
To allow for more accurate predictions as shown in the literature review, we will enrich our dataset with additional exogenous factors. The occupancy data of the buildings are unavailable but outdoors weather conditions can be obtained from nearby weather stations. Unfortunately, the dataset used by the previous student became unavailable, but a new source was found that can supply weather data from a nearby station.

The weather dataset consists of many factors that could possibly affect the energy load predictions. These will need to be assessed in future work of how they relate to the output of forecasting models and select the important ones for increasing accuracy in the model.

3 Conclusion and future work

The analysis of building data and creation of energy forecasting models brings great benefits environmentally and economically. Literature review has shown multiple cases were using energy consumption forecasting models, accurate predictions can be made that result in increased energy efficiency. The main focus of the work presented in this paper was replicating previous work that can be reused for creating an energy forecasting system. This work has presented numerous energy load forecasting models that need to be assessed in future work and completed all preliminary steps. We have a clean dataset with basic knowledge about the data and additional exogenous data for enriching our dataset thus allowing us to dive into the exploration of time series forecasting models. After assessing multiple methods and choosing the proper model we will try to make the process interpretable by exploring interpretability techniques. This will help validating our model by giving us a deeper insight into how it works. After achieving this, future work will focus on implementing the forecasting model into a system that could process streaming data from the sensors obtain exogenous factors from weather stations automatically and detect outliers in real time.

4 Project management



Analysis and optimisation of building management systems through data analytics and interpretable machine learning

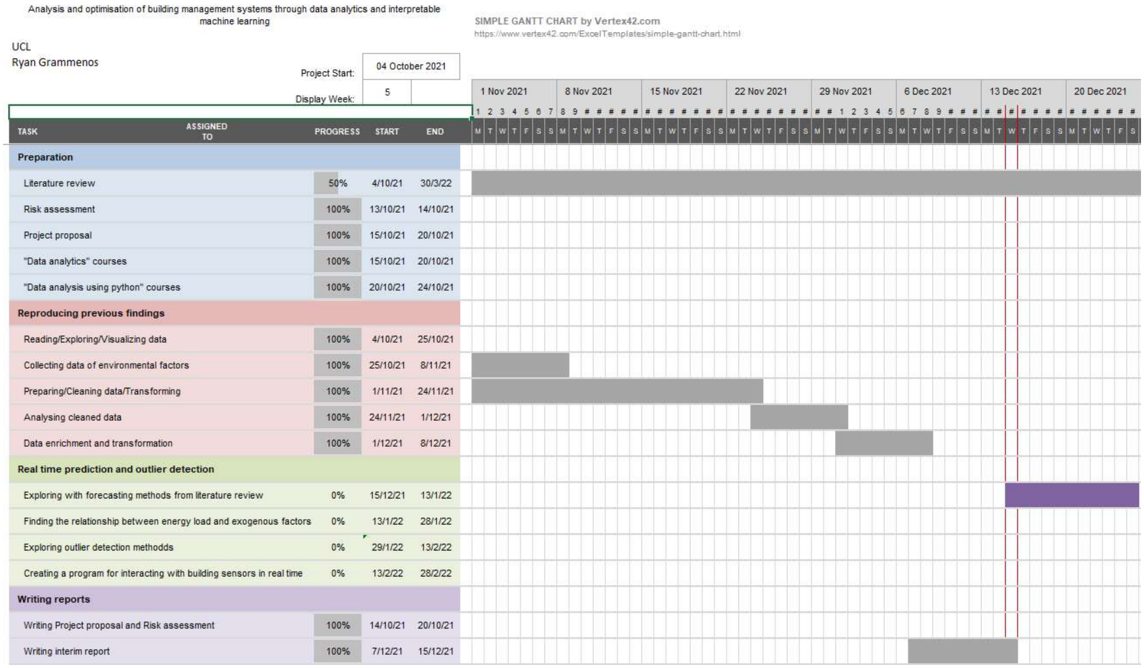
SIMPLE GANTT CHART by Vertex42.com
https://www.vertex42.com/ExcelTemplates/simple-gantt-chart.html

UCL
Ryan Grammenos

Project Start: 04 October 2021
Display Week: 1

TASK	ASSIGNED TO	PROGRESS	START	END
Preparation				
Literature review		50%	4/10/21	30/3/22
Risk assessment		100%	13/10/21	14/10/21
Project proposal		100%	15/10/21	20/10/21
"Data analytics" courses		100%	15/10/21	20/10/21
"Data analysis using python" courses		100%	20/10/21	24/10/21
Reproducing previous findings				
Reading/Exploring/Visualizing data		100%	4/10/21	25/10/21
Collecting data of environmental factors		100%	25/10/21	8/11/21
Preparing/Cleaning data/Transforming		100%	1/11/21	24/11/21
Analysing cleaned data		100%	24/11/21	1/12/21
Data enrichment and transformation		100%	1/12/21	8/12/21
Real time prediction and outlier detection				
Exploring with forecasting methods from literature review		0%	1/12/21	8/12/21
Finding the relationship between energy load and exogenous		0%	8/12/21	23/12/21
Exploring outlier detection methodds		0%	24/12/21	8/1/22
Creating a program for interacting with building sensors in real time		0%	8/1/22	23/1/22
Writing reports				
Writing Project proposal and Risk assessment		100%	14/10/21	20/10/21
Writing interim report		100%	7/12/21	15/12/21

Figure 15 Progress in first term - Part 1



Analysis and optimisation of building management systems through data analytics and interpretable machine learning

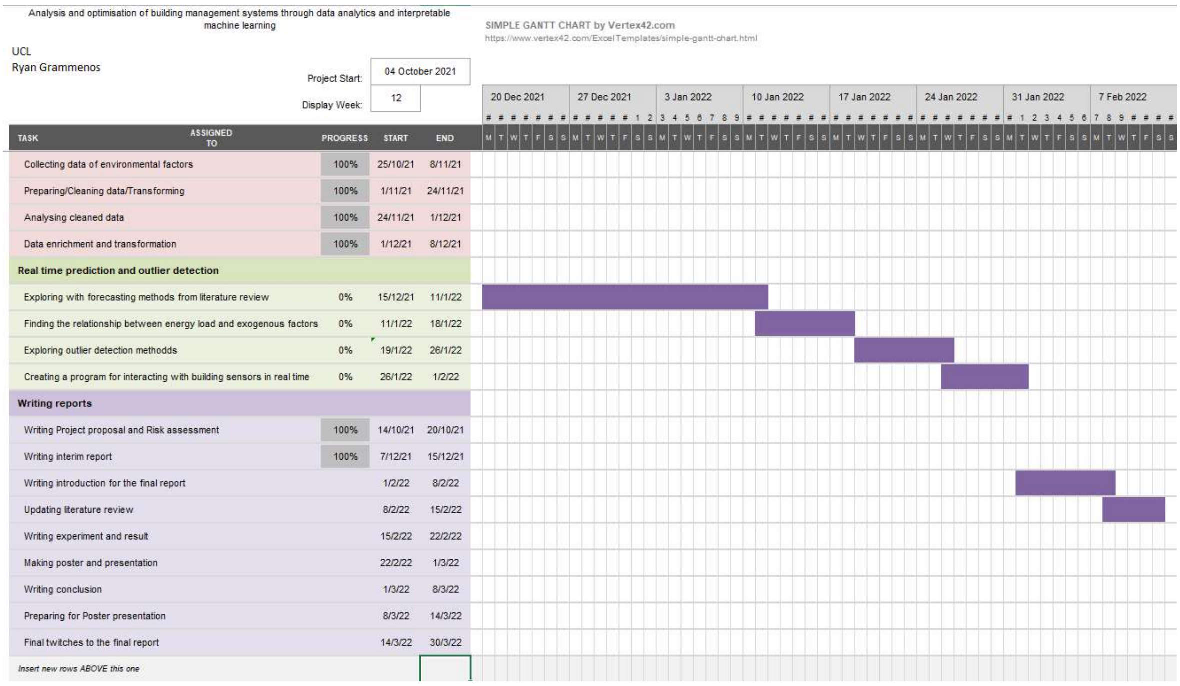
SIMPLE GANTT CHART by Vertex42.com
https://www.vertex42.com/ExcelTemplates/simple-gantt-chart.html

UCL
Ryan Grammenos

Project Start: 04 October 2021
Display Week: 5

TASK	ASSIGNED TO	PROGRESS	START	END
Preparation				
Literature review		50%	4/10/21	30/3/22
Risk assessment		100%	13/10/21	14/10/21
Project proposal		100%	15/10/21	20/10/21
"Data analytics" courses		100%	15/10/21	20/10/21
"Data analysis using python" courses		100%	20/10/21	24/10/21
Reproducing previous findings				
Reading/Exploring/Visualizing data		100%	4/10/21	25/10/21
Collecting data of environmental factors		100%	25/10/21	8/11/21
Preparing/Cleaning data/Transforming		100%	1/11/21	24/11/21
Analysing cleaned data		100%	24/11/21	1/12/21
Data enrichment and transformation		100%	1/12/21	8/12/21
Real time prediction and outlier detection				
Exploring with forecasting methods from literature review		0%	15/12/21	13/1/22
Finding the relationship between energy load and exogenous factors		0%	13/1/22	28/1/22
Exploring outlier detection methodds		0%	29/1/22	13/2/22
Creating a program for interacting with building sensors in real time		0%	13/2/22	28/2/22
Writing reports				
Writing Project proposal and Risk assessment		100%	14/10/21	20/10/21
Writing interim report		100%	7/12/21	15/12/21

Figure 16-Progress in first term - Part 2



Analysis and optimisation of building management systems through data analytics and interpretable machine learning

UCL
Ryan Grammenos

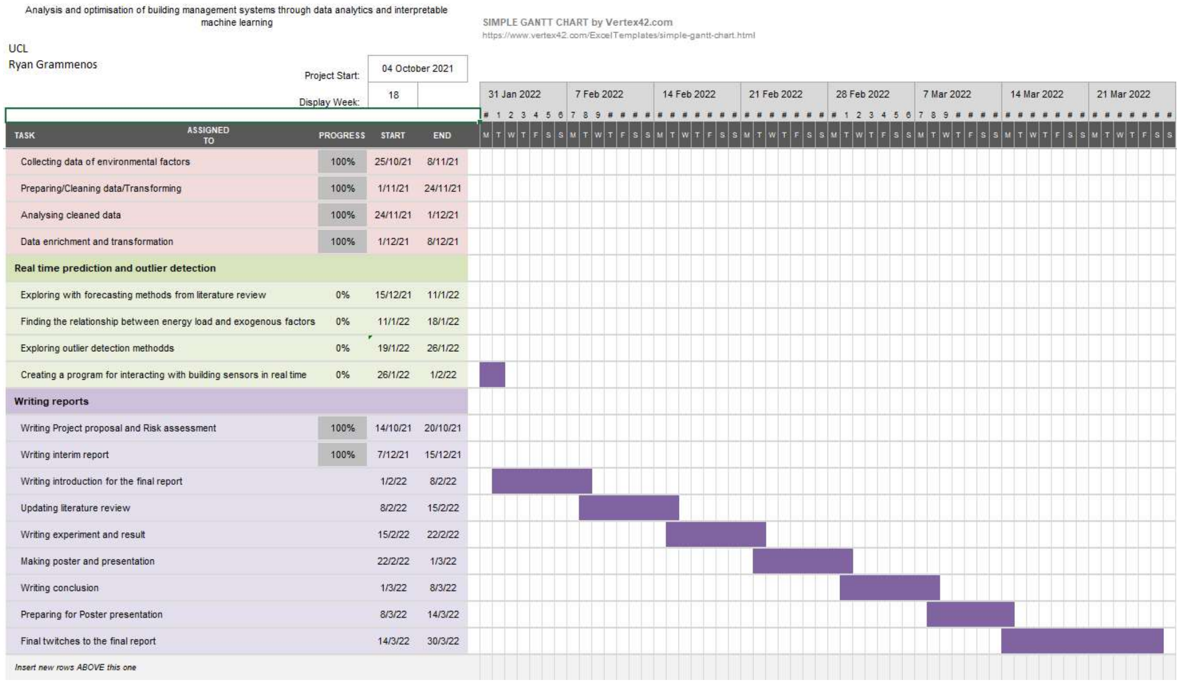
Project Start: 04 October 2021

Display Week: 12

SIMPLE GANTT CHART by Vertex42.com
https://www.vertex42.com/ExcelTemplates/simple-gantt-chart.html

TASK	ASSIGNED TO	PROGRESS	START	END
Collecting data of environmental factors		100%	25/10/21	8/11/21
Preparing/Cleaning data/Transforming		100%	1/11/21	24/11/21
Analysing cleaned data		100%	24/11/21	1/12/21
Data enrichment and transformation		100%	1/12/21	8/12/21
Real time prediction and outlier detection				
Exploring with forecasting methods from literature review		0%	15/12/21	11/1/22
Finding the relationship between energy load and exogenous factors		0%	11/1/22	18/1/22
Exploring outlier detection methodds		0%	19/1/22	26/1/22
Creating a program for interacting with building sensors in real time		0%	26/1/22	1/2/22
Writing reports				
Writing Project proposal and Risk assessment		100%	14/10/21	20/10/21
Writing interim report		100%	7/12/21	15/12/21
Writing introduction for the final report			1/2/22	8/2/22
Updating literature review			8/2/22	15/2/22
Writing experiment and result			15/2/22	22/2/22
Making poster and presentation			22/2/22	1/3/22
Writing conclusion			1/3/22	8/3/22
Preparing for Poster presentation			8/3/22	14/3/22
Final twitches to the final report			14/3/22	30/3/22
Insert new rows ABOVE this one				

Figure 17 - Action plan for Term 2 - Part 1



Analysis and optimisation of building management systems through data analytics and interpretable machine learning

UCL
Ryan Grammenos

Project Start: 04 October 2021

Display Week: 18

SIMPLE GANTT CHART by Vertex42.com
https://www.vertex42.com/ExcelTemplates/simple-gantt-chart.html

TASK	ASSIGNED TO	PROGRESS	START	END
Collecting data of environmental factors		100%	25/10/21	8/11/21
Preparing/Cleaning data/Transforming		100%	1/11/21	24/11/21
Analysing cleaned data		100%	24/11/21	1/12/21
Data enrichment and transformation		100%	1/12/21	8/12/21
Real time prediction and outlier detection				
Exploring with forecasting methods from literature review		0%	15/12/21	11/1/22
Finding the relationship between energy load and exogenous factors		0%	11/1/22	18/1/22
Exploring outlier detection methodds		0%	19/1/22	26/1/22
Creating a program for interacting with building sensors in real time		0%	26/1/22	1/2/22
Writing reports				
Writing Project proposal and Risk assessment		100%	14/10/21	20/10/21
Writing interim report		100%	7/12/21	15/12/21
Writing introduction for the final report			1/2/22	8/2/22
Updating literature review			8/2/22	15/2/22
Writing experiment and result			15/2/22	22/2/22
Making poster and presentation			22/2/22	1/3/22
Writing conclusion			1/3/22	8/3/22
Preparing for Poster presentation			8/3/22	14/3/22
Final twitches to the final report			14/3/22	30/3/22
Insert new rows ABOVE this one				

Figure 18 - Action plan for Term 2 - Part 2

Figure 15 and 16 show the progress over term 1. Comparing to initial plans I did not achieve my goals and I am behind on my work. Changes were necessary due to unforeseen problems such as spending additional time on understanding the reason for the differences in the results for the cleaning algorithm when using the flowchart code whilst other plots show a correct cleaning algorithm. A further problem was not allocating time for writing the risk assessment, project proposal and interim report. The workload resulting from these was worth at least two weeks of work (30+ hours.). Overall, the replication of the previous study was finished thus allowing me to move on to creating unique work individually. The action plan for term 2 can be seen on figures 17 and 18. A significant amount of time was allocated to exploring forecasting methods due to the immense workload and the timing of the holiday season. After selecting the forecasting model, I hope to finish remaining tasks on a weekly basis. To avoid problems with report writing, a clear outline of the dates for finishing parts of the report has been provided. Time was allocated generously to allowing me to tackle unforeseen difficulties.

5 References

[1] F. M. Hai-xiang Zhao, “A review on the prediction of building energy consumption,” *Renewable and Sustainable Energy Reviews,* vol. 16, no. 6, pp. 3586-3592, 2012.

[2] I. &. C. A. &. C. S. &. C. T. Khan, “Fault Detection Analysis of Building Energy Consumption Using Data Mining Techniques,” *Energy Procedia,* vol. 42, pp. 557-566, 2013.

[3] A. Z. A.-G. SSAK Javeed Nizami, “Forecasting electric energy consumption using neural networks,” *Energy Policy,* vol. 23, no. 12, pp. 1097-1104, 1995.

[4] M. S. V. G. S. Karatasou, “Modeling and predicting building's energy use with artificial neural networks: Methods and results,” *Energy and Buildings,* vol. 38, no. 8, pp. 949-958, 2006.

[5] G. R. a. B. B. J. Newsham, “Building-Level Occupancy Data to Improve ARIMA-Based Electricity Use Forecasts,” in *Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building*, New York, NY, USA, 2010.

[6] L. J.-h. D. L. X. G. a. L. J. L. Xuemei, “Building Cooling Load Forecasting Model Based on LS-SVM,” in *2009 Asia-Pacific Conference on Information Processing*, vol. 1, Shenzhen, China, IEEE, 2009, pp. 55-58.

[7] Z. L. H. Z. P. X. Yangyang Fu, “Using Support Vector Machine to Predict Next Day Electricity Load of Public Buildings with Sub-metering Devices,” *Procedia Engineering,* vol. 121, pp. 1016-1022, 2015.

[8] Y. C. ,. X. S. ,. D. W. Junhua Zhuang, “Building Cooling Load Prediction Based on Time Series Method and Neural Networks,” *International Journal of Grid and Distributed Computing,* vol. 8, no. 4, pp. 105-114, 2015.

[9] G. L. X. L. Y. W. Hongzhan Nie, “Hybrid of ARIMA and SVMs for Short-Term Load Forecasting,” *Energy Procedia,* vol. 16, no. C, pp. 1455-1460, 2012.

[10] K. a. S. W. a. S. N. a. M. X. Yu, “Real-time Outlier Detection Over Streaming Data,” in *2019 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, IEEE, 2019, pp. 125-132.

[11] S.-l. a. S. C.-x. a. Z. Q. a. Z. X.-x. Lei, “2005 IEEE Russia Power Tech,” in *The research of local linear model of short term electrical load on multivariate time series*, St. Petersburg, Russia, IEEE, 2005, pp. 1-5.