Waterford Institute of Technology

APPLYING DATA MINING TECHNIQUES TO MONITORING DATA FROM A DATA CENTRE TO IMPROVE ITS ENERGY EFFICIENCY

Research Proposal

Student
John Fitzpatrick,
02323826

Supervisor
Dr. Bernard Butler

December 18, 2018

Contents

List of Figures

Listings

1. Introduction

The demand for cloud computing services is growing exponentially, driven by the growth of internet users and applications (Mahdavi and Tschudi, 2013). As a result, within the Information and Communications Technology sector (ICT), Data Centres are the fastest growing sector in terms of energy consumption and are now responsible for over 2% of the global CO2 emissions (Avgerinou, Bertoldi, and Castellazi, 2017), with this growth in Data Centre energy consumption over the last 10 years, there arises a cost incentive, along with a general green incentive, to reduce energy consumption or to use energy more efficiently. While there is ample literature available that deals with making Data Centres more efficient by changing the hardware configuration of the Data Centre, this proposal will focus on applying Data Mining techniques to the monitoring data, compiled from within a Data Centre, to unlock hidden information to assist system administrators to make the Data Centre more efficient. Data from a medium sized Data Centre, the WIT-TSSG Data Centre in the Carriganore West Campus of Waterford institute of Technology is used for this proposed research.

1.1. Motivation

The goal behind making Data Centres more energy efficient is to reduce the energy consumption while delivering the same service to its users. Therefore the motivation that follows from this is twofold.

Firstly, there is the motivation to make the Data Centre more energy efficient for the tangible benefits this brings. If energy consumption is reduced, this will lead to cost savings, while simultaneously reducing the carbon footprint of the Data Centre. Ultimately, this could deliver savings to WIT, allowing for potential investment in other educational projects while helping the environment at the same time.

Secondly, in addition to the overarching motivation above, this proposal is motivated by a desire to explore Data Mining techniques to gain insight into Data Centre energy consumption to help make them more energy efficient. While Data Centres are a part component of any cloud computing architecture, the emphasis here is to treat them as an entity in their own right, and use the vast amounts of monitoring data they generate to make them more energy efficient. There is very little existing literature in this regard and the motivation here is that this study would contribute to the research in this field.

1.2. Research Problem Statement

As outlined in the introduction (section 1), The information and Communication Technology (ICT) sector, and Data Centres in particular, are consuming vast amounts of energy and are contributing to an ever increasing carbon footprint. Avgerinou, Bertoldi, and Castellazi (2017) highlight that the ICT sector will consume 13% of the global electricity supply by 2030.

Therefore, as the energy demand of Data Centres grows due to the changing role of technology in society, this represents a clear challenge to policymakers, researchers and

Data Centre stakeholders to seek solutions to make Data Centres more efficient in terms of energy consumption.

A significant amount of research exists that focuses on making Data Centres more efficient, by concentrating on improvements to the physical hardware and infrastructure within a Data Centre. This, as detailed in the literature review (section 2.3), specifically focused on such measures as:

1. Introducing wireless sensors to capture key metrics;
2. Following best practices and
3. Fcussing on specific features within a Data Centre to gain efficiencies, such as the cooling system, server efficiencies, network modelling, or the optimisation of applications at an infrastructural level.

However, this research has a different focus. It will analyse existing monitoring data, that is already stored by the Data Centre to seek hidden insights that might make a Data Centre more energy efficient.

1.3. Aims and Objectives

The ultimate aim of the research proposed here, as specified in the Research Problem Statement (section 1.2) is to use Data Mining Techniques on the monitoring data compiled at the WIT-TSSG Data Centre to gain hidden insights about energy consumption to help make the Data Centre more energy efficient. Specifically, the following are the aims and objectives of the project:

- Review all existing literature on improving Data Centre energy efficiency.
- Learn about the WIT-TSSG Data Centre and what monitoring data is compiled that can be used for analysis.
- Identify Data Mining techniques from existing literature that can be applied to the monitoring data compiled at the Data Centre.
- Apply Data Mining techniques, which are still to be decided, at this stage of the proposal, on the monitoring data from the Data Centre to maybe gain some hidden insights about energy consumption.

If successful with these aims, the objective is to understand and predict when there is a peak in energy demand so that the supply of energy can be better utilized. If the project can meet this objective, it will also have contributed to the academic literature on this subject, while simultaneously providing the Data Centre in WIT-TSSG with a predictive model to help better utilize the supply of energy at peak demand that in turn can lead to greater energy efficiencies.

1.4. Research Questions

From monitoring data provided by the WIT-TSSG Data Centre and using a Data Mining based approach,

1. Can we identify predictive indicators from the data that forecast an unanticipated peak in energy demand at the Data Centre?

2. Can we identify features in the data that indicate where the Data Centre is not optimally consuming its supply of energy?

2. Preliminary Literary Review

2.1. Introduction

The literature reviewed at this point can be divided into three concepts. The first is an introduction to what Data Centres are, how they work and specifically an overview of the WIT-TSSG Data Centre in Carriganore in Waterford (section 2.2).

The next concept looks at the literature around improving the energy efficiency of Data Centres (section 2.3). In most cases, the focus is on reducing power consumption by improving the efficiency of different hardware components of a Data Centre. Specifically, the research looks at either the server, networking or cooling systems with an emphasis, by multiple authors, on implementing a wireless sensor system to capture key metrics across a Data Centre and using this to pinpoint issues in real-time.

Other work in this area presents research on optimising resources or developing new scorecards to identify efficiencies in the management of Data Centres.

The final concept looks at research on Data Mining techniques used to increase energy efficiency in buildings (section 2.4). There are numerous case studies using 'Big Data' from building automation systems, to be harnessed in a Data Mining approach to identify hidden insights to improve the energy efficiency of different buildings. All three concepts are explored in detail below, before a summary ties all three concepts together and identifies where any research gap is.

2.2. Data Centre Infrastructure and the WIT-TSSG Data Centre

2.2.1. What is a Data Centre and How Much Energy do they Use?

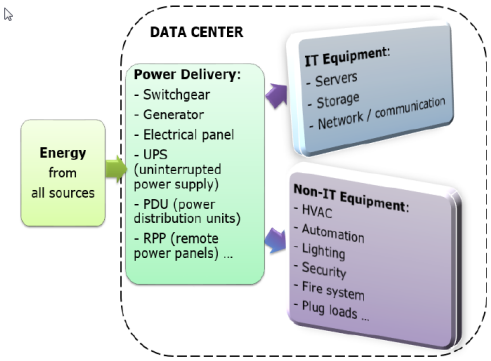
In the modern economy, the demand for cloud storage and cloud computing based applications is growing exponentially as the number of internet users who rely on digital information via cloud services continues to sky-rocket across the world ((Vasudevan et al., 2017)). To facilitate the demand in growth for such services, it follows that Data Centres, and the facilities they provide, will also be more in demand.

Levy and Raviv (2018) define a Data Centre as "a dedicated facility with all the resources required for storage, processing and sharing digital information and its supported

areas". The hardware required for such a facility comprises power distribution, environmental control systems, telecommunications, security and information technology (I.T) equipment (Levy and Raviv, 2018).

Figure 1 from (Levy and Raviv, 2018) outlines how energy is typically consumed in a Data Centre, specifically how it is channelled to IT equipment and non IT equipment.

Figure 1: How Energy is Consumed in a Data Centre.



Data Centres consume forty times more energy than conventional office buildings (Levy and Raviv, 2018) and have the fastest growing carbon footprint across the whole ICT sector (Avgerinou, Bertoldi, and Castellazi, 2017).

Figure 2, reproduced from Avgerinou, Bertoldi, and Castellazi (2017) shows the rising energy consumption of Data Centres across the EU and the US over the last 15 years as the number of Data Centres have grown:

Figure 2: Energy Consumption of Data Centres in the EU and US from 2000 to currently.

Consumption (TWh)	Reporting Year
	EU consumption
18.3	2000
41.3	2005
56	2007
72.5	2010
104	2020
	US consumption
91	2013
140	2020
	Global consumption
216	2007
269	2012

With Data Centres consuming more energy, there is an obvious incentive, both from a cost and a green perspective, to reduce energy consumption per Data Centre. This can be done by making Data Centres more efficient for the energy they consume and the measurement that is used to measure energy efficiency is the *Power Usage Efficiency* (PUE) score. This is described as

Power Usage Effectiveness = Total Facility Energy Usage/ IT Equipment Usage.

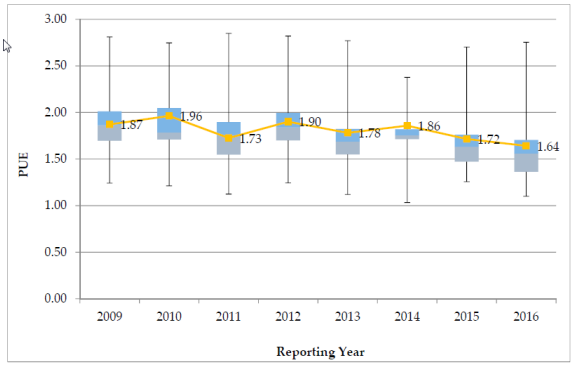
Lower PUE scores are better, and a PUE close to one indicates that the Data Centre is more efficient as it consumes less energy on non IT equipment such as cooling. In terms of the current trends of energy usage, research by Avgerinou, Bertoldi, and Castellazi (2017) show that Data Centres in Northern Europe have a lower PUE score than those in southern Europe, which is expected given the geographical and climate difference. Also, the same authors find that the PUE scores in general have been decreasing year after year as demonstrated by the graphs in Figures 3 and 4.

Figure 3: The average PUE across different locations.

Table 3. Geographical zoning with temperature and relative humidity average data.

Geographical Zones	Countries	Temperature Range (°C)	RH Range (%)	Average PUE	Number of Data Centres
Nordic countries	Denmark, Finland, Norway, Sweden	18–26	20–80	1.71	13
UK and Republic of Ireland	England, Scotland, Wales, Northern Ireland, Republic of Ireland.	17–30	8–80	1.83	116
Northern/Central Europe	Austria, Belgium, France, Germany, Hungary, Luxembourg, The Netherlands, Portugal, Poland, Switzerland	14–28	16–75	1.72	122
Southern Europe/ Mediterranean	Gibraltar, Greece, Italy, Malta, Spain, Turkey, Monaco, Romania, Bulgaria	16–26	20–80	2.00	30
Non EU	Republic of Mauritius, US	-	-	-	5

Figure 4: The average PUE from 2000 to currently.



So the challenge, as outlined in the various literature below (section 2.3), is how to make a Data Centre more energy efficient by maximizing the energy used for IT equipment and reducing the consumption of energy by non-IT equipment.

2.2.2. The WIT-TSSG Data Centre

The WIT-TSSG Data Centre is located on the Carriganore West Campus of the Waterford Institute of Technology (WIT) on the west side of Waterford city. From this location, the Data Centre supports over 50 concurrently active ICT research projects through provisioning of internet services, Cloud Computing resources, an AI cluster and project bespoke test beds (*www.tssg.org* 2018).

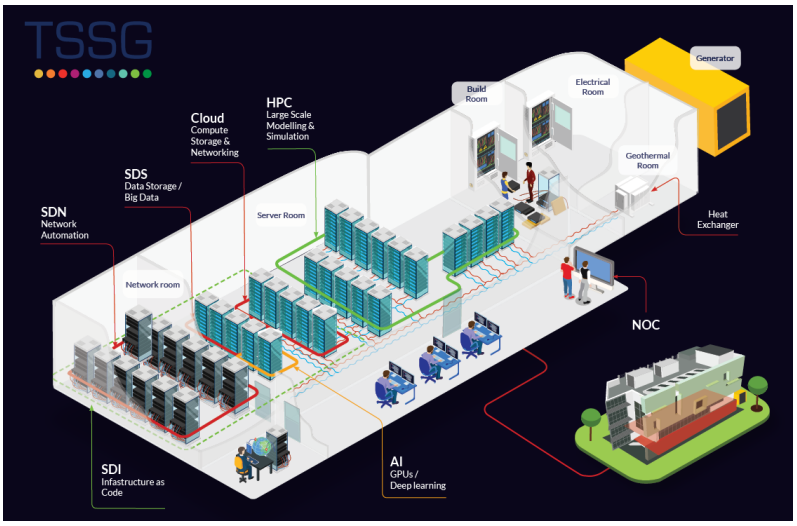
In particular, the Data Centre supports TSSG projects, the Higher Education Authority, The Irish Centre for High End Computing(ICHEC), hosting its super computer Kay, Arclabs and other 3rd party commercial arrangements.

In terms of its energy usage, the Data Centre has an IT power load of 300kW, with cabinets engineered to house 30kW of IT equipment. 1+1 300kW UPSs and an 800kW Generator provide the backup power (*www.tssg.org* 2018).

With regards to physical hardware, the Data Centre supports over 160 physical servers, providing more than 1,000 cpu cores for processing and 400 virtual servers for cloud computing. In addition, there is over half a PB (that is 512 TB) of Data Storage, and 3,000 network port (*www.tssg.org* 2018).

Figure 5, provided from *www.tssg.org* (2018) demonstrates the design and infrastructure of the Data Centre building.

Figure 5: The design of the WIT-TSSG Data Centre.



From this diagram we see that there are 26 Cabinets and these use varying degrees of KW per cabinet to house over 160 servers. In addition, the Data Centre is designed so that there are separate rooms for servers, networking equipment, UPS and batteries with the generator and cooling equipment outside. This is to optomise the best cooling option for each room, resulting in the server room not having any hot or cold aisles. Instead, all heat and cooling is contained inside the 26 cabinets (*www.tssg.org* 2018).

In terms of a cooling system, the Data Centre uses a water based cooling system which uses less energy than an air cooled system. For monitoring and visualisation, the Data Centre uses Nagios and the ELK software stack (**ELK**) with the output from that monitoring system is discussed further in section 4.1 and 4.2 (Data Available and Exploratory Data Analysis).

2.3. Improving Data Centre Energy Efficiency

From the Research Problem Statement (section 1.2) the benefits for improving Data Centre energy efficiency is twofold: reducing the carbon footprint of the Data Centre, and reducing the cost of maintaining the Data Centre and these benefits should be mutually inclusive. Therefore, with such a desired outcome at stake, it is not surprising

that a large amount of literature devotes itself to efforts in improving energy efficiency in Data Centres.

As outlined from the previous section (2.2), Data Centres are recent innovations, so to a degree, the literature is still catching up with the newest technologies and best practices. However, it is important to outline the earlier literature to understand why modern Data Centres, such as the WIT-TSSG Data Centre would have the ability to capture and store sensory data from its general operations. The following research summarises the need for such a sensor system and the immediate benefits this can bring to the energy efficiency of a Data Centre.

To begin, a 2008 study by Masanet et al. (2011) states that implementing a sensor system to capture data can reduce energy demand by as much as 80% through aggressive pursuit of energy efficiency measures specifically focussed around servers, while in 2012, the literature from this time also calls for the implementation of wireless sensors to begin capturing data.

Here, Mahdavi and Tschudi (2013) argues that capturing this data via wireless sensors and then using integrated software to manage and integrate this into management systems can allow for greater efficiencies in a Data Centres cooling system and power usage, while Liu and Terzis (2012) explain that a lack of visibility into a Data Centres operating conditions is an underlying reason for low-energy utilization.

In this study, the authors argue for the introduction of sensor networks that combined with modelling, could provide insights into the distribution of energy and resolve the tension between cooling and performance.

While these outlined the immediate and obvious benefits of integrating more reliable data into management systems and decisions, which by and large has been implemented, more recent literature investigates where more marginal gains can be found.

Rygielski, Kounev, and Tran-Gia (2016) looks to the performance of network capabilities and in particular, this research develops simulation models to predict the performance of networks based on the traffic through networks and concludes that the network performance models only lead to minor prediction accuracy degradation but the same models can accelerate the performance analysis by a much larger factor.

Arjona et al. (2014) uses sensor data and a Map-Reduce computation in Hadoop to analyse server usage to maximise the efficiency of CPU power being used by each server in a Data Centre, thereby introducing some advanced data analysis of a specifically gathered sensor data.

Marcos-Jorquera et al. (2016) moved the argument on from creating sensor networks to developing a monitoring service for energy efficiency and sustainable management in Data Centres. The approach the authors use is to collect the monitoring data, analyse the data captured and then execute based on the outcome of the analysis. This is further developed today in Data Centres where centres use the TICK or ELK (section 2.2.2).

Another approach used to improve Data Centre energy efficiency in the literature is advanced by Levy and Raviv (2018) who introduce a whole new set of metrics to review the efficiency of Data Centres. They argue that while the old metrics are valuable, some key metrics concerning *risk* are overlooked. They advance a new Data Centre multidimensional scorecard which gives better visualisation to identify areas of improvement.

More advanced initiatives are based around the optimisation of physical resources within a Data Centre. Bi et al. (2015) propose a multi-tier architecture-orientated virtualised application performance model to allocate computing resources to each machine based on SLA (service level agreement) restrictions. The authors develop a heuristic mixed optimisation algorithm to achieve this.

Finally, Vasudevan et al. (2017) seek to reduce the operational costs of Data Centres and maximise energy efficiency by analysing log data to profile applications using the Data Centre. In addition, the authors profile the virtual and physical machines in the Data Centre, and using a penalty-based matching algorithm match different applications with the most compatible machines (virtual and physical) so that CPU usage is more efficient.

The literature outlined here demonstrates the changing focus of how best to increase the energy efficiency of data centres. From early calls, just ten years ago, to implement sensor systems to that would lead to improved management systems and easily identified gains, to more recent and more advanced techniques looking to use algorithms to increase optimisation of the resources in a Data Centre. In all cases highlighted, the importance of capturing, understanding and using data to integrate new information into management systems and operations is stressed. However, in these instances, the approaches identify explicit data they want to capture and analyse, whether this is log data on CPU utilization or specific monitoring data. In general, few Data Mining approaches are used, in the sense of using implicit data to uncover hidden insights. The research questions in section 1.4 focuses on this concept.

2.4. Data Mining Techniques (for improving energy efficiency in buildings)

Data Mining is the process of uncovering hidden patterns from large sets of data, where the data is analysed to extract information which can be transformed into valuable knowledge for the future. Such techniques include statistical and machine learning techniques which take th form of regression analysis, classification, cluster analysis and more advanced techniques such as neural networks. These techniques analyse the relationship between variables to uncover such hidden patterns and trends.

While there is limited literature or existing research on using Data Mining techniques to improve the energy efficiency of Data Centres, there is comprehensive research on using Data Mining approaches to assess the energy efficiency of buildings in general, or of various systems within buildings. Because a Data Centre is a building in its own right, it is important to review such Data Mining techniques in the building management literature and assess if they can be applied to Data Centres.

In general, the literature can be divided into two sections, where basic Data Mining techniques are used, as described above, and when more complex algorithms are used to gain insights. First, I explore the literature on the basic Data Mining approaches.

Xue et al. (2017) look at district heating substations in Northern China, and from operational data collected from the substations use clustering techniques to identify

seasonal and daily operating patterns, and association rules mining to identify faults, malfunctions and inefficient operations strategies.

Kuo, Lin, and Lee (2018) uses regression and classification techniques on different factors in analysing the energy consumption characteristics of Taiwan's convenience stores. The data is collected from a survey of convenience stores owners before the Data Mining techniques look at intercorrelation between energy consumption and other factors such as geography, climate and store design.

Using operational data from the construction of an education building in Hong Kong University, Fan et al. (2018) use decision trees, motif discovery and gradual pattern mining to uncover novel and valuable insights for the management of energy consumption for the building.

Also, using monitoring data from building automation systems, Fan et al. (2018) explore how unsupervised Data Mining techniques such as clustering, association rule mining, motif discovery and unsupervised anomaly detection can be used to gain insights into energy efficiencies within buildings.

In a similar study by Zhou et al. (2018), the authors use data supplied from heat supply companies in Singapore and look at the correlation between energy consumption and building physics, the heating system and room position, and use the information gain ratio algorithm and decision tree classifiers to analyse energy consumption and other related factors.

Ahmad and Chen (2018) use Gaussian process regression, multiple linear regression, tree bagger, bagged tree, boosted tree and neural networks to identify abnormal behaviour and predict the future heating and cooling demands in building environments.

However, some research also uses more advanced algorithms to gain hidden insights from the data captured.

Sönmez et al. (2018) use log data and the self-organising maps algorithm to create a decision support application for anomaly detection in IT systems. They highlight that this algorithm is used in anomaly detection in fraud cases and security attacks, but most importantly they highlight, when using the log data from systems, the self-organising maps algorithm allows for early detection of failures in critical systems in real-time.

The Random Forrest Algorithm is used by Wang et al. (2017) to discover critical events for event driven optimisation in building air-conditioned systems. With data collected from the building automation system, the authors make a case for using Data Mining to gain efficiencies in air conditioned systems by moving away from a time based system, to a more event driven optimisation.

The literature reviewed highlights the wide and varied Data Mining approaches taken by various authors to gain hidden insights into energy efficiencies within buildings or building systems such as heating or cooling systems. From this, I will outline below how I propose we can adapt these approaches to the data we have gathered from WIT-TSSG Data Centre to help answer our Research Questions (section 1.4).

2.5. Summary

While the literature reviewed here is preliminary and further literary research will continue in the next stage of the project, the research questions above (section 1.4) clearly set out the goal of maximising energy efficiency within a Data Centre and the preliminary literature review sets out the framework for building the foundations for this research. Firstly, it defines what a Data Centre is and the trends in energy consumption by Data Centres (section 2.2.1) before introducing the WIT-TSSG Data centre in Waterford (section 2.2.2). From these foundations, the literature review focussed on various computational and systematic improvements that could lead to greater energy efficiency within a Data Centre (section 2.3). The importance of wireless sensory data is a common starting point for most authors who focus on efficiency improvements to be made by concentrating on hardware improvements (network, server or cooling systems). Most of these approaches use data which is specifically captured at a system level to assist the goal of increased energy efficiency.

However, in all this research, there is no instance, that I could find, of an author trying to find hidden insights from any operational or monitoring data already gathered from a Data Centre for another purpose. Some authors mention that this is the next step or as an area that should be looked at as a future study, but in the main, the concept of increasing the energy efficiency of Data Centres in the literature, concentrates on the establishment of monitoring systems to capture data for a specific end goal.

This is where I believe there is a research gap. Data Centres are entities in their own right and store vast amounts of operational and monitoring data. This has been harvested by a monitoring system and could be useful to analyse, using Data Mining techniques, to find hidden insights around energy efficiency.

Having discovered a research gap, the final section of the literary review looked at where Data Mining techniques are used for similar applications (section 2.4) such as the energy efficiency of different building types, or heating or cooling systems within a building. Various Data Mining techniques are explored in the literature and are assessed to see if such approaches can be adapted for use in a Data Centre.

3. Working Theory

From the literature reviewed here (section 2), a gap was identified where data mining techniques could be used on the monitoring data compiled within a Data Centre to unlock hidden sources of information that could lead to greater energy efficiency in operating the Data Centre. The research questions (section 1.4), ask if these efficiencies could brought about by being able to predict when energy demand peaks in order to better utilise the supply of energy that the Data Centre is using.

Whereas the majority of literature in this area analyses the hardware components of a Data Centre to improve efficiency (section 2.3), this research proposes to concentrate on the energy demand of a Data Centre. At any given time, a Data Centre will have a certain load to process, and will utilise it's available resources to process the load. To

maximise efficiency therefore, a Data Centre would need to be able to best match the energy demand with the supply at its disposal. The real gains in efficiency therefore, is to be able to anticipate the demand to best utilise the resources at its disposal. In technical terms, this would mean moving loads between racks within a Data Centre or reducing the demand for the cooling system at certain times.

Therefore, to increase the efficiency of a Data Centre by concentrating on energy demand, we need to anticipate the demand of energy usage. This proposal seeks to create such a predictive model using the Data Mining techniques outlined in the literature review (section 2.4) and the monitoring data collected from within the Data Centre (section 4.1).

If a good predictive model of energy demand can be built, this will be the missing piece of the jigsaw in helping the Data Centre become more efficient by analysing its own monitoring data to anticipate a surge in energy demand. This in turn will allow the Data Centre to utilise its supply of energy making it more energy efficient and aligning with our aim of reducing the cost of running the Data Centre (section 1.3)

In conclusion, this proposal will treat a Data Centre as an entity in its own right. Instead of merely being an end point in a cloud computing infrastructure for many applications or physical machines, it will utilise the data created by the Data Centre itself. By doing this, a predictive model can be built to anticipate the demand of energy needed, thereby transforming the Data Centre from being an intricate part a cloud computing framework to a more Edge orientated framework in its own right.

4. Research Methods

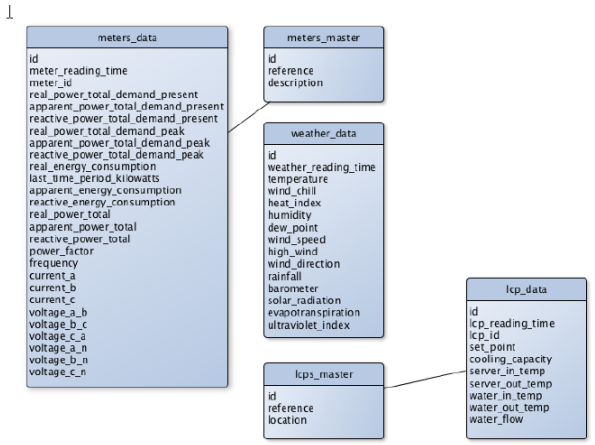
This section builds on the Working Theory and firstly introduces the data available to this study before delving further into how a predictive model might be built by using the data available.

4.1. Data Available

As outlined in section 2.2.2, the Data Centre examined is the WIT-TSSG (Telecommunications, Software and Systems Group) Data Centre in the Carriganore West Campus, part of the Waterford Institute of Technology (WIT). The Data Centre was completed in January 2010 in order to support its network-based research projects in the area of telecommunications networking and cloud computing. The Data Centre was designed with high-density, power efficiency and separation of services as its primary goals. It has a total IT power load of 300Kw with some cabinets engineered to house 30kw of IT equipment.

Similar to most modern Data Centres, sensors and the use of network devices are used to monitor and capture data on how the Data Centre is operating. As outlined in the schema model shown in Figure 6 below, there is substantive qualitative data captured that can be used for the empirical investigations to create a predictive model.

Figure 6: The schema of the WIT-TSSG Data Centre.



As we see from the schema, the data captures a multitude of metered energy usage variables, different weather data, and temperature and water usage data. The data outlined here is available from 2014 right up to present, captured at 5 minute intervals, making this a time series data set. In the next section, the data is explored at an initial level to find any seasonal trends or predictive variables.

4.2. Exploratory Data Analysis

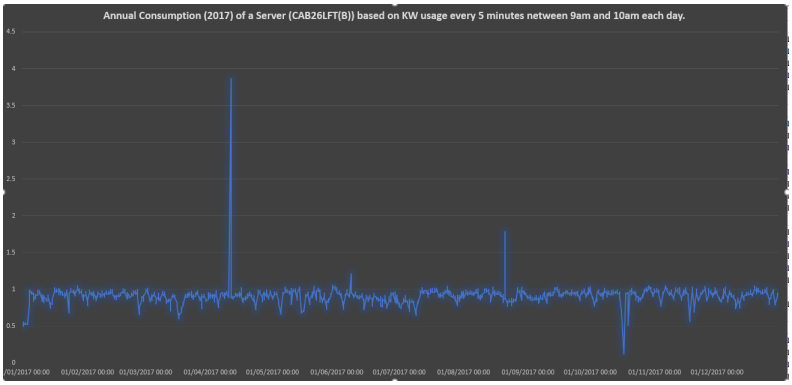
Having briefly introduced the data, it is also important to note that this is a static 'data dump'. Additional data, when available, can be added to the study if required. The data was accessed by connecting remotely into the Data Centre via PuTTY and querying the data via MySQL (See Appendices A). From here, the CSV output from the SQL queries was SFTP'd to my local machine and exported into Microsoft Excel where some preliminary data analysis was carried out.

The schema introduced in Figure 6 outlines the vast scale of variables available and as such data is gathered in 5 minute intervals over four years, the amount of data accumulated is vast and can be challenging to work with. While it is proposed that Python will be used for Data Mining Techniques to investigate the data as outlined in section 4.3, Line charts are ideal to illustrate what the time series data can show and the following charts are snippets from the data to demonstrate the range of data available.

Figure 7, generated using Listing 1, illustrates the energy consumption of one particular server cabinet within the Data Centre (Cabinet 26A) over the course of 2017. The daily readings are taken from between 9am and 10am each day and while the chart may initially suggest that the energy usage has been relatively consistent, there are significant spikes in energy consumption in April and August, with some considerable drops in demand too. The Research Techniques section (4.3) will outline how this study would seek to investigate if these spikes can be predicted so that another cabinet could share some of the load to reduce energy consumption within the Data Centre.

However, while Figure 7 was a snapshot of energy usage over a year, the data can also

Figure 7: Energy consumption (in kW usage per 5 minute intervals) of one cabinet in the Data Centre over the entire year of 2017.



be presented for a particular day.

Figure 8: Energy consumption (in kW usage per 5 minute intervals) of three different server cabinets in the Data Centre over the course of one day (2017-09-13).

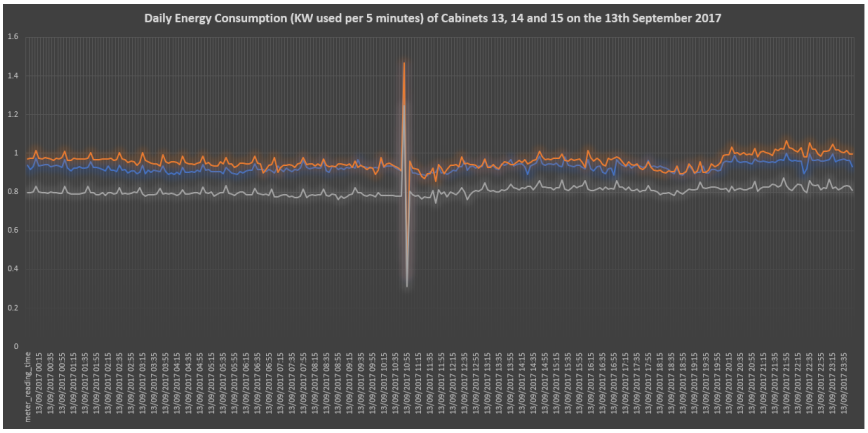
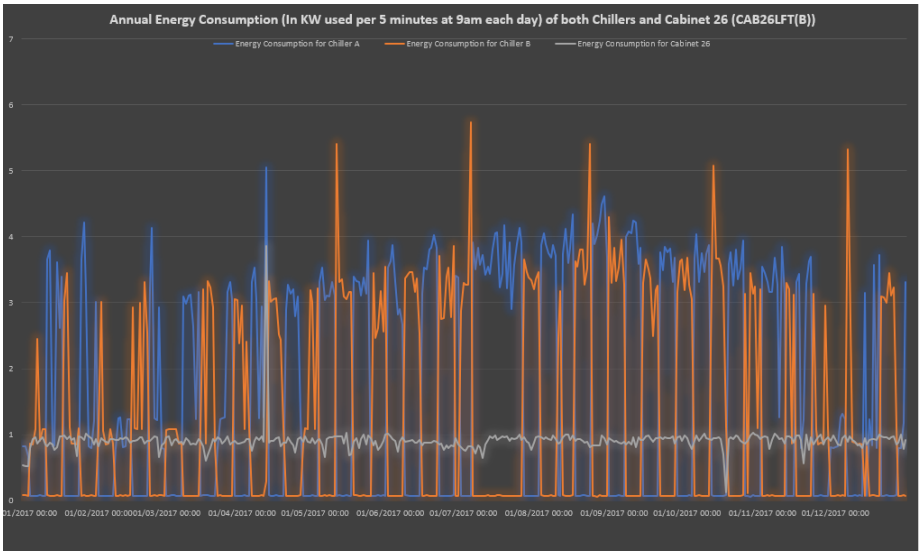


Figure 8 reports significant spikes in energy consumption in two of the cabinets around 11am. Again, this is another example of how Data Mining techniques might be applied to assist in being able to predict when such spikes occur.

While certain spikes in energy consumption might be suitable for further investigation, the preliminary data analysis should be able to detect some seasonal behaviour. Figure 9 demonstrates this. While the energy usage for the server cabinet (cabinet 26b) is relatively stable throughout the year, the energy consumption for the chillers shows a clear increase throughout the warmer summer months. Interestingly however, while each chiller works on alternate weeks, there are unexplained spikes for chiller B throughout the second half of the year that are not replicated when chiller A is running.

Other interesting preliminary analysis of the data centred around water consumption at the Data Centre (Figure 10) and the energy usage of the generator during a severe weather event, in this case being Storm Ophelia on October 2017 (Figure 11).

Figure 9: Energy consumption (in kW usage per 5 minute intervals) of both Chillers and one server cabinet in the Data Centre over the course of 2017



All these charts are provided to simply demonstrate the range of data available. The next section discusses how this data could be screened to build a predictive model and the challenges that arise from this.

4.3. Research Techniques

An advantage for this proposed research is that the data required is already available for analysis and the previous sections (4.1 and 4.2) introduced the variables and did some exploratory data analysis. This section progresses further, and proposes how the data might be screened to build a predictive model as outlined in the Working Theory.

Figure's 7, 8, 9, 10 and 11 demonstrate that the data available here is time series data. With data going back to 2014, this allows for investigations into the energy usage per cabinet, per rack or even at a system level, whether that is the energy consumed by the network configuration, the server demand or other computational needs.

This study proposes to screen these variables, and discover predictive variables of when energy demand peaks. The Exploratory Data Analysis section (4.2) even highlighted instances of when demand actually peaks. Predicting this would be helpful as it will allow the system administrators to manage the supply of energy better at these peak times, thereby increasing the efficiency of the data centre. The Exploratory Data Analysis section (4.2) also illustrates seasonal effects where energy demand will increase in the summer months of the year as the warm weather precipitates the need for greater cooling within a Data Centre. However, the approach here will be to analyse these factors at a deeper level to build a more developed predictive model that unlocks hidden insights not yet discovered.

It is planned that for the next stage of the project, the data will be imported into

Figure 10: Annual water flow (in litres per 5 minute intervals) to LCP 1 measured at 9am every day during 2017

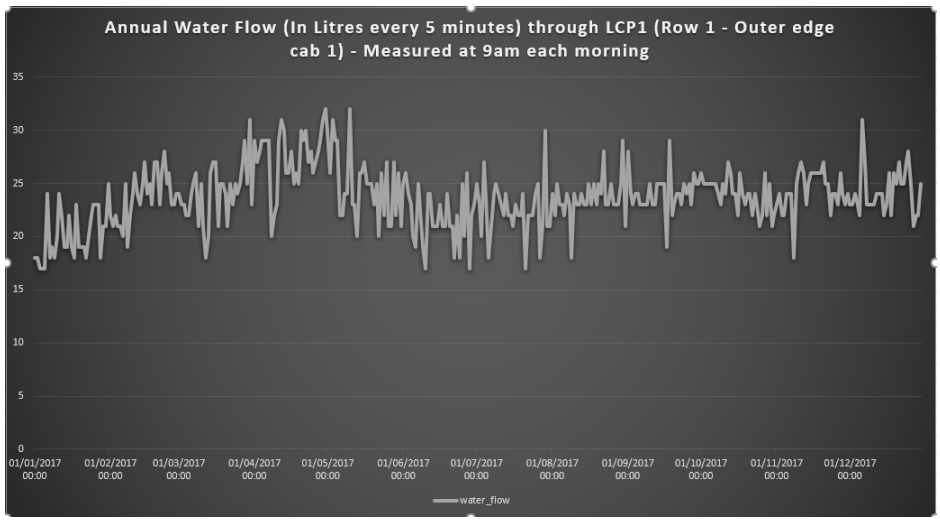
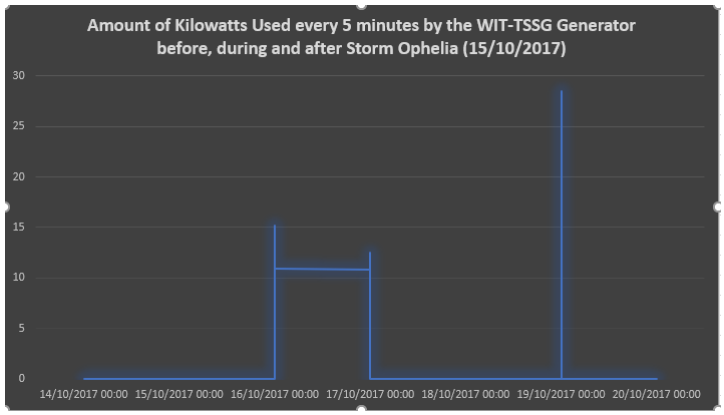


Figure 11: Energy consumption (in kW usage per 5 minute intervals) of the Generator before, during and after Storm Ophelia in October 15th 2017



Python for detailed Data Mining exploration. The literature review above (section 2.4) mentioned various Data Mining techniques that have been used to increase energy efficiency of different building and heating systems. Some of these will be utilised here, and various regression and clustering analysis techniques will be used in Python to identify predictive variables and build a predictive model. Regression will look at the relationship between variables and predict a quantitative output, so potentially this could predict energy consumption by looking at different servers combined with weather information or the supply of water to the chiller systems. The objective would be to find ways of deriving suitable features to improve the model predictions. Clustering techniques could be used to look at the relationships between these predictors and derive some intuition regarding correlations between different events and problems with the chillers. These are just some examples that will be explored but ultimately at this stage

of a project, the exact method is not known yet. There will be a degree of trial and error in finding the exact method and the most suitable data.

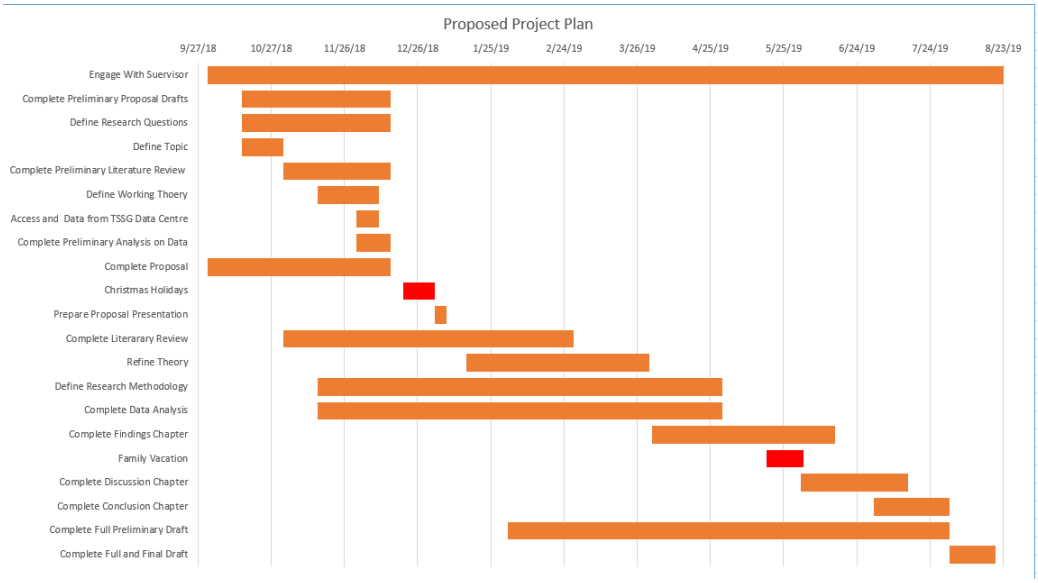
However, a predictive model is only useful if it's predictable and reliable, meaning it should be accurate and precise. In attempting to develop such a model it will be important to focus on goodness of fit measures like bias and variance. This will be a key factor in the project also.

Finally, as outlined in section 4.1, it is important to note that the data presented here is a static 'data dump'. This allows for the development of predictive models on the static data only. However, an advantage of having nearly five years of static data is that it allows for the model to be tested on different subsections of the data to test its validity and reliability. While this is useful and would solve for our research questions (section 1.4) concerning increasing energy efficiency within a Data Centre, a more ideal model would be based on live streaming data. This would allow us to solve for problems outlined in real-time and develop models that will assist in anomaly detection or highlight when system failures might occur in real time and assist in avoiding such failures.

5. Project Plan

The proposed plan of work for this project is outlined in the Gantt chart in Figure 12 below. The sections in red is when I am considered not available. The remainder of the tasks have added contingency to avoid a chaotic and disordered critical path and align with the key milestone dates for Thesis submission as outlined by WIT authorities.

Figure 12: Project Plan for 2018-2019.



6. Proposal Summary

This proposal began by outlining the growing prevalence that Data Centres play in the ICT sector and how much energy they currently consume. This set the scene for understanding the incentives and benefits of increasing the energy efficiency within Data Centres. The literature review then introduced Data Centres in more detail, including the WIT-TSSG Data Centre, before surveying the existing literature that focused on increasing the energy efficiency of Data Centres. The literature review also looked at how Data Mining Techniques was used in increasing energy efficiency of other building types before a gap was established where Data Mining techniques could be used to help Data Centres become more energy efficient.

The working theory section expanded on this gap and set the framework for the research methods section which firstly explained the data available before looking at how the project might proceed, using Data Mining techniques and the data at hand to develop a predictive model that will predict energy demand. This proposal concludes with a prospective project plan for how this research would proceed in the new year.

References

Ahmad, Tanveer and Huanxin Chen (2018). "Short and medium-term forecasting of cooling and heating load demand in building environment with data-mining based approaches". In: *Energy and Buildings* 166, pp. 460–476. ISSN: 0378-7788.

Arjona, Jordi et al. (2014). "A Measurement-based Analysis of the Energy Consumption of Data Center Servers." In: *CoRR* abs/1402.0804. ISSN: 14020804.

Avgerinou, Maria, Paolo Bertoldi, and Luca Castellazi (Oct. 2017). "Trends in Data Centre Energy Consumption under the European Code of Conduct for Data Centre Energy Efficiency." In: *Energies* 10.1470. ISSN: 1996-1073.

Bi, Jing et al. (2015). "SLA-based optimisation of virtualised resource for multi-tier web applications in cloud data centres." In: *Enterprise Information Systems* 9.7, pp. 743–767. ISSN: 17517575.

Fan, Cheng et al. (2018). "Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review". In: *Energy and Buildings* 159, pp. 296–308. ISSN: 0378-7788.

Kuo, Chung-Feng Jeffrey, Chieh-Hung Lin, and Ming-Hao Lee (2018). "Analyze the energy consumption characteristics and affecting factors of Taiwan's convenience stores-using the big data mining approach". In: *Energy and Buildings* 168, pp. 120–136. ISSN: 0378-7788.

Levy, Moses and Daniel Raviv (2018). "An Overview of Data Center Metrics and a Novel Approach for a New Family of Metrics." In: *Advances in Science, Technology and Engineering Systems, Vol 3, Iss 2, Pp 238-251 (2018)* 2, p. 238. ISSN: 2415-6698.

Liu, Jie and Andreas Terzis (2012). "Sensing data centres for energy efficiency." In: *Philosophical Transactions: Mathematical, Physical and Engineering Sciences* 1958, p. 136. ISSN: 1364503X.

Mahdavi, Rod and William Tschudi (2013). "Wireless Sensor Network for Improving the Energy Efficiency of Data Centers." In: *Lawrence Berkely National Library*.

Marcos-Jorquera, Diego et al. (Sept. 2016). "Smart Monitoring Embedded Service for Energy-Efficient and Sustainable Management in Data Centers." In: *Energies* 9.515. ISSN: 9070515.

Estimating the Energy Use and Efficiency Potential of U.S. Data Centers. (2011). Proceedings of the IEEE.

Rygielski, Piotr, Samuel Kounev, and Phuoc Tran-Gia (2016). "Flexible Performance Prediction of Data Center Networks using Automatically Generated Simulation Models." In: *EAI Endorsed Transactions on Scalable Information Systems, Vol 3, Iss 9, Pp 1-10 (2016)* 9, p. 1. ISSN: 2032-9407.

Sönmez, Ferdi et al. (2018). "Anomaly Detection Using Data Mining Methods in IT Systems: A Decision Support Application." In: *Sakarya University Journal of Science* 22.4, p. 1109. ISSN: 13014048.

Vasudevan, Meera et al. (2017). "Profile-based application assignment for greener and more energy-efficient data centers". In: *Future Generation Computer Systems* 67, pp. 94–108. ISSN: 0167-739X.

Wang, Junqi et al. (2017). "A Data Mining Approach to Discover Critical Events for Event-Driven Optimization in Building Air Conditioning Systems". In: *Energy Procedia* 143. Leveraging Energy Technologies and Policy Options for Low Carbon Cities, pp. 251–257. ISSN: 1876-6102.

www.tssg.org (2018). `http://www.tssg.org/research/research-supports/infrastructure-group-2/`. Accessed: 2018-12-12.

Xue, Puning et al. (2017). "Fault detection and operation optimization in district heating substations based on data mining techniques". In: *Applied Energy* 205, pp. 926–940. ISSN: 0306-2619.

Zhou, Hao et al. (2018). "Analysis of correlation between actual heating energy consumption and building physics, heating system, and room position using data mining approach". In: *Energy and Buildings* 166, pp. 73–82. ISSN: 0378-7788.

Appendices

A. MySQL Queries for Exploratory Data Analysis

This section provides the MySQL queries used to source the data for the line charts in Figures 7, 8, 9, 10 and 11.

Listing 1: Annual Energy Consumption

```
use dc;
select *
from meters_data
where meter_reading_time > '2017-01-01'
and meter_reading_time < '2017-12-31'
and HOUR(`meter_reading_time`) = 9
and meter_id not in ('1','2','3','4','5','6','7','8','9','10','11',
 '12','77','78');
```

Listing 2: Daily Energy Consumption

```
use dc;
select *
from meters_data
where meter_reading_time > '2017-09-12'
and meter_reading_time < '2017-09-14'
where meter_id in ('13','14','15');
```

Listing 3: Annual Energy Consumption of Chillers and Server

```
use dc;
select *
from meters_data
where meter_reading_time > '2017-01-01'
and meter_reading_time < '2017-12-31'
and HOUR(`meter_reading_time`) = 9
and MINUTE(`meter_reading_time`) = 0
and meter_id in ('3','4','13');
```

Listing 4: Annual Water Flow

```
use dc;
select *
from lcps_data
where lcp_reading_time > '2017-01-01'
and lcp_reading_time < '2017-12-31'
and HOUR(`lcp_reading_time`) = 9
and MINUTE(`lcp_reading_time`) = 0;
```

Listing 5: Energy Consumption of the Generator during Storm Ophelia

```
use dc;
select *
from meters_data
where meter_reading_time > '2017-10-10'
and meter_reading_time < '2017-10-24'
and meter_id = '2';
```