Data-driven approaches to creating optimal energy consumption policies for multi-function print devices

MComp Computer Science with Industrial Placement

Supervisor: Matthew Forshaw

May 2020

Callum Matthew Simpson – 16030326

Abstract

This dissertation investigates if the use of data analysis could be used to create a more optimal solution to a real-world problem. The problem discussed in this paper will be the current energy-saving policy tied to a device with the planned solution being a policy that balances energy-performance trade-off. Different areas will be analyzed to discover what aspect of a device polices should be tailored too.

Declaration

I declare that this dissertation represents my work except where otherwise stated.

Acknowledgments

I would like to give a big thank you to my dissertation supervisor Dr Matthew Forshaw for planting the seeds for this dissertation. He has helped the organization of this project and has provided me with a lot of helpful advice and resources whenever I have needed it.

Table of Contents

Chapter 1: Introduction

1.1 Subject area

There are many electronic devices that are set up ready to be used. At the time of writing this, there are an estimated 26.66 billion IoT devices installed worldwide and it's predicted that this number will increase to around 75 billion by 2025 (Statista, 2020).

Currently, most UK households contain around 50 such devices (Energysavingsecrets, 2020). Whilst every device usage differs, some device will never be turned off when not in use. This can lead to standby power consumption making up to 22% of a device's total energy consumption (Francis, 2014).

The issue of power wastage has already been investigated and regulations have already been passed to combat it, for example, EC Regulation No. 1275/2008 which means that standby power for electrical and electronic household and office equipment shall not exceed 0.5W (Commission Regulation, 2008)

Even with regulations established to reduce waste per device, when many of these devices are grouped the amount wasted starts to add up. UK businesses are spending £60m a year in avoidable electricity bills, which equates to enough electricity to power 65,000 homes and producing carbon emission to match 46,000 cars (Shrestha, 2020).

To combat this, most devices have an energy consumption policy put in place that helps by switching the device to an ideal energy-saving state when not in use. However, most polices are programmed for static use and do not account for device high and low usage peaks meaning they will be sub-optimal at certain times. Printers are an example of one these devices and when printing is 10-16% of ICT related energy consumption in higher education (James, 2009), the small amount of waste per device will add up.

This paper seeks to investigate if there is a better way to handle energy consumption policies or any way to enhance currently existing policies using information retrieved via data analysis to reduce the amount of energy used and the number of delays to a user.

1.2 Aims and objectives

1.2.1 Aim

My project aims to investigate how data analyses can be employed to create an optimal energy consumption policy that balances energy-performance trade-off in a collection of printers placed in different clusters in hopes of uncovering methods that could be employed on other devices. This would save energy, time and money. Though there has been work, looking into the ways we can optimize energy policies not many of them use data analysis to help them and most of them study devices like CPU's.

After investigating current policies that are in use, I will simulate each policy against real data to discover what aspects of a device usage a policy should be tailored to. I will only be looking at reducing energy wastage by adjusting the energy policy and not by modifying the hardware or any of the other software/features.

The most optimal policy will be the one that produces a low power consumption and narrows the number of delays to the user (aka finding a policy that has the best amount of uptime and downtime). Finding this balance is important as if a user wants to use the printer but the printer is in a sleep state then energy will be used converting it to a state that can print. This transition uses more energy than switching from idle to printing. If a user finds that they always need to turn a printer on, then it would damage the printer in the long run and would wastage more energy than leaving it idle.

1.2.2 Objectives

1: Research and summarize state of the art energy policies

There has been a lot of work gone into designing different policies that combat energy consumption. If I'm to create a policy I must gain an understanding of what is currently out there and how they work. I will be conducting background reading to accumulate papers that will assist me in achieving my goals.

2: Perform data analyses to find trends in printer usage

Using a set of anonymized data of print jobs done by a set of printers located over three cluster rooms and user logins session for those cluster rooms, I am planning to analyses the data to get a better understanding of realistic printer usage. Using statistical analysis techniques (or other techniques if appropriate), I will be looking for trends and peaks in the printer usage.

3: Implement and simulate different policies

A simulation will be developed that can run a policy against a set of print jobs. The simulation will allow me to gather information about how effective each policy is. Simulated policies will be either preexisting policies or ones created using data analysis findings.

4: Modify policies to discover potential avenues for optimization

Once a simulation has been carried out, I will analyze the results to see if I can find areas that I could try to tweak to try to increase policy efficiencies.

5: Analyze and compare different policies to find the optimal energy-performance trade-off

Once I have collected all my results that I have gotten throughout my investigation I will compare them to each other to find out which one is the most optimal. Again, the optimal value will be the one that finds the best balance between low power consumption and causes limited delays.

My findings here will allow me to discuss what changes we could make to current energy policies so that they become more optimal.

Chapter 2: Project methods

2.1 Technology

Data cleanup was done using Microsoft Excel and Microsoft PowerBI (Microsoft, 2018). Excel provides many helpful tools that can be used to aid data analyses especially in cleaning data as changes have an immediate visual impact. Another benefit is allowing for quick formatting of cells. Power BI is a collection of software used in business analysis that I primarily used it for easy merging of data sets.

Most of my actual analysis was done in R using R studio (R studio, 2020) with the packages TidyVerse (TidyVerse, 2020) and Ggplot2 (Ggplot2, 2020). R is an interpreted language widely used by statisticians and data miners to do data analysis as it provides a wide variety of features for statistical techniques. TidyVerse is a collection of open-source packages designed for data science, increasing the number of features to aid with data analysis. Ggplot2 was chosen to help with the visualization of data as it has been designed with the deep philosophy of visualization in mind.

Originally the simulation was going to be made using an in-house High-throughput computing simulation (Forshaw, 2019) that would have been modified for this dissertation. The simulation was a Java-based and trace-driven, which was originally developed to study energy-saving policies. However, due to unforeseen setbacks, I ended up in a position where I needed to create a new java simulation from scratch.

2.2 Approach

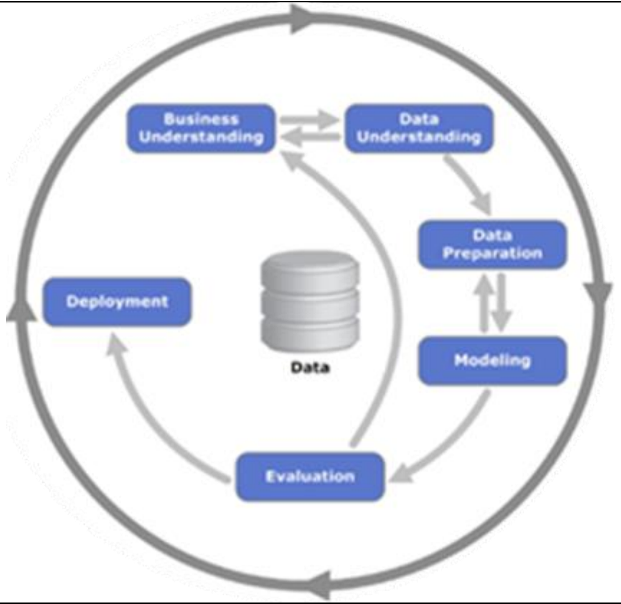


Figure 1 - The CRISP-DM model (Jensen, 2012)

I plan is to complete my dissertation by following the CRISP-DM model. The Cross-industry standard process for data mining (*Chapman, 2019)* is one of the most widely used analytical models that it used in industry. This is because the CRISP-DM is a cycle model where the sequence of steps is not rigid and what I do next will depend on what my exploratory data analysis tells me. The CRISP-DM will allow me to loop back and try different things. This is good for my project as each iteration will provide me with a new set of data that will allow me to do further tests to find the most optimal energy policy.

2.2.1 Business understanding

The Business-understanding step focuses on the creation and understanding the aims and objectives of the given project at hand. I have set it so this step is where I do all my research but that may change as I might do more research throughout my first few cycles. This will be working towards my first objective.

2.2.2 Data understanding

The data-understanding phase is where I will be reading through the data I have been given to identify any data quality problems and ultimately obtain an understanding of the data that I have been given to work out ways to phrase possible questions.

2.2.3 Data preparation

Data preparation will see me cleaning data and constructing a dataset that I will be preforming my data analysis on.

2.2.3 Modelling phase

I will be modelling the different policies that I have come across. It possible that one of my policies has a specific data requirement, if so; I will go back to the data preparation step to get the needed data. I will then test each of the policies in a simulation of realistic printer usage to see how they hold up. The simulation will be created to read in a history of print jobs and printers. Different policies created and current standards will be implemented into this simulation.

2.2.4 Evaluation phase

In the evaluation step, I will evaluate my findings. To do a full evaluation I will need to go back to the business-understanding step to make sure my results fully match the requirements that I set. If the results have not met the requirements or if I feel like I could refine them in some way, then I will loop through every step again to get a set of results that better meets my aim. Each loop through this will push me towards my fifth objective and the data I have achieved in this step will be used to help decided where I will go next, linking to my fourth objective.

2.2.5 Deployment step

For my dissertation, the deployment step will be writing up my findings.

2.3 Problems

This project was my first time doing anything serious with data science so there was a lot for me to learn which took up time. I had also never used R before but with help from my supervisor and guides online, I was able to get a decent grasp on the language.

The worldwide spread of COVID caused a few issues for this dissertation. Originally, this dissertation was going to use an in-house High-throughput computing simulation that was created to calculate energy saving that was going to be modified to suit this dissertation. Though this would have been useful to my dissertation, it would have taken a lot of work and effort from both myself and my supervisor to get it working. I decided it would be best to go ahead and build a basic version of the simulation so that I could generate some results. This took time to get working correctly. It also meant that some of the analysis into user logins had to be scrapped as given the limited time period it would have been difficult to get it working correctly.

I also ran into a few problems with the data, but I will discuss that in Chapter 4.2.

2.4 Ethics and data

Resnik (Resnik, 2000) states that it’s unethical to fail to disclose important information relevant to the data analysis, such as any assumptions made about the parameters or methods used. This is because a researcher should make every effort to be upfront and unbiased about data. (Kassner, 2017) Failure to provide honest and accurate insights may cause massive issues on the results due to misguided data. Not only does this help reduce errors it could also be a type of misconduct if done with big data (Shamoo and Resnik, 2014).

To ensure that I meet these standards I have asked myself question posed in the AREA 4P Framework (Orbit RRI, 2020) to help me ensure the acceptability and desirability of my research.

- I believe the purpose of this research to be beneficial as it could help produce a method that could be used to help reduce energy wastage in other devices.
- I decided that I would be using statistical analysis as I want to find what there is causing the peaks device usage. This will be covered in Chapter 5.1.
- The data I have received has been anonymized meaning I could not retrieve any data that a person might feel is unsafe. Each print job has a username that does not match actual records. I also cannot see what the person has printed off.
- I will not edit individual values to try a skew to the outcome, the only changes I may make on the data is changing names of things to make it easy to follow or removing data in which is an error (clear duplication or corrupted values).
- Any interpretation that I have about my findings will clearly be stated.
- All analysis will be done with outliers included.

Chapter 3: Research

3.1 Data analysis

As my dissertation heavily relies on data analysis, I thought it would be beneficial to discuss what exactly is data analysis to explain the steps I have followed to complete my dissertation.

3.1.1 What is data analysis

Data analysis can be defined as a process of collecting, cleaning, transforming and then modelling data to discover useful information that could help make a decision. Shamoo and Resnik (Shamoo and Resnik, 2006) state that data analysis "provides a way of drawing inductive inferences from data and distinguishing any real phenomena or effects from random fluctuations."

In a sense when we think about our experiences to help us influence our future actions, self-reflection to make decisions, then that is what data analysis is all about.

3.1.2 The steps of data analysis

Data analysis is done over several steps. In the following section, I will be going through the steps involved with data analysis in the order that they are normally completed and will discuss why they are done.

3.1.2.1 Data Requirement Gathering

Data analysis starts with requirement gathering. This phase starts with asking yourself what you want to do analyses on and what you are overall aiming to find.

"No amount of statistical analysis, regardless of the level of the sophistication, will correct poorly defined objective outcome measurements. Whether done unintentionally or by design, this practice increases the likelihood of clouding the interpretation of findings, thus potentially misleading readers." (RCR, 2020).

More importantly, a decision should be made on what type of analysis wants to carry out and how results are going to be measured. Once the question has been created, a set of objectives should also be outlined to help guide analysis. There are many different analytical techniques that a person can decide to use depending on what they want to find, though the form of analysis is usually determined by the specific data gathering method chosen. According to Smeeton and Goda (2003), "Statistical advice should be obtained at the stage of initial planning of an investigation so that, for example, the method of sampling and design of questionnaire are appropriate".

3.1.2.1.1 Text analysis

Transform raw data into information by spotting patterns in a set of data. This is normally done using databases or data mining tools.

3.1.2.1.2 Descriptive Analysis

Analyses on complete data to find out "what happened". For continuous data, it can be used to find mean and deviations whereas on categorical data it can be used to find percentage and frequency.

3.1.2.1.3 Inferential Analysis

Using the same data set, split the data into subsets and analysis each of them to see if each provide different conclusions.

3.1.2.1.4 Diagnostic Analysis

Finding the "why something happens" by searching for behavior patterns to gain insight into what the caused a problem. This is typically done alongside descriptive analysis because descriptive will show that something has taken place meaning diagnostic can be used to figure out the reason behind it. If the reason for why a problem happened, then that issue can be prevented from happening again.

3.1.2.1.5 Predictive Analysis

Finding the "what is likely going to happen". Making predictions about the future based on what has happened in the past. The accuracy is based on how much detailed information is collected and how in-depth the analysis is done.

3.1.2.1.6 Prescriptive Analysis

Applying a wide range of analytical techniques to work out what should be done to solve the current problem. By using, a range of techniques then more information may be found but they may not be as in-depth as just using one method and it may take more time.

3.1.2.2 Data collection

The next step is collecting the data based on the requirements that where set out previously. Data can be either quantitative data (numerical information) or qualitative data (wordier/descriptive response). The sources data can be gathered from are primary collection and secondary.

3.1.2.2.1 Primary collection

This is data that the researcher themselves has collected for their research, typically in the form of polls, surveys, interviews or statistical groups. This data is unique and is normally tailored for the researcher's problem. It is more precise but could be time-consuming to collect.

3.1.2.2.2 Secondary data

Data collected by another researcher for their own research that has already been passed through some sort of analysis. This normally comes in the form of reports and statements. This data is quicker to gather but it may not be fully tailored to help the problem at hand.

It's important that a plan is created to make sure that only relevant information is collected and any person identifiable data collected for one person cannot be used for any means that aren't strictly tied to original purpose without the consent of the examined person (Harriss,2018).

3.1.2.3 Data cleaning

Data cleaning is arguably the most important part of data analysis. This revolves around scrubbing out errors, missing values, duplicates and any value that violates the current rules by removing or replacing the value. It's important that data is clean as leaving in errors may result in an incorrect conclusion or have a real business impact. According to "The average financial impact of poor data quality on organizations is $9.7 million per year" (Fernandez, 2019). This loss could be due to the finical costs, as poor data costs a business at least 30% of their revenue, missed market opportunities or damages to a brand image as miss-communication due to bad data could lead to negative branding.

3.1.2.3.1 Requirements for clean data

To ensure that data is cleaned to the highest quality the following criteria should be met (Wikipedia Data ,2020)

3.1.2.3.1.1 Validity

The values must conform to the constraints of the column. Possible examples of constraints are datatype (I.e a column for Booleans should only contain Booleans), range and uniqueness. Validity must hold across multiple columns as well, cross-field validation is checking that conditions that span across multiple fields hold, for example, a user cannot be recorded as leaving a building before their recoded entering. Validity test is simple, search a column to see if all rows meet a condition, failed rows stand out. For example, check employee IDs to make sure that no two employees share the same code.

3.1.2.3.1.2 Accuracy

The level of cleanliness at which the data can be called true or realistic. Accuracy is not the same as validity as you can have a value that meets the constraints of the column but does not accurately represent the truth. The "perfect dataset" is often unachievable because discovering what values are true values and which are false but valid can only be done by using an external source that contains information for comparison. For example, is hard to tell if some has given their real name without looking it up.

3.1.2.3.1.3 Consistency

In the dataset, there are no values that contradict each other. Consistency is also another problem that is difficult to solve perfectly. If two values that contradict each other, say a user is recorded as having 2 different addresses, then it's difficult to work out which one which is the real one without having to backtrack.

3.1.2.3.1.4 Completeness

The data contains enough values to be able to meet all requirements of the analysis. Data that is deemed incomplete cannot normally be fixed by cleaning because the information missing may contain key facts that could be beneficial to the analysis. Data that is deemed incomplete can only be fixed by gathering more data.

3.1.2.3.2 Techniques to clean data

To meet these requirements many different techniques could be applied, each having its pros and cons. All techniques effectively try to reduce the amount of incorrect data by removing, correcting or imputing it. (DigitalVyda , 2018).

One quick way to clean the data is by removing columns or rows (based on a column values) that contain information that do not fit the context of the problem that is being solved. However, a column that may seem irrelevant may contain an interesting correlation with another column that could be interesting to analyze.

Another is searching for duplicate data points that are repeated throughout the data. Duplicate rows show that something has happened in the gathering processes that has caused a duplicate to happen. This is simple to resolve by removing one duplicate so that one remains. However, if duplication is found that breaks the validity of uniqueness then further investigation should be involved to determine which one should be kept, or which one should be modified.

Type conversion can be done to a full column to convert all values to the same unit measurement. If there were values that could not be converted, I.e. a string to a number, then the conversion would produce an NA values indicating that something has gone wrong but this allows a user to see which values would need extra work to get them to the correct format.

By doing a count of a column, not only would this show NA values, it would also show if any of the values contain a typo or any syntax issues as those with an issue would be represented as a new class.

Missing or NA data can be dealt with in three ways, simply remove the rows that contain a NA value, flagging the missing data if the NA don't appear to be caused by random chance (i.e. people with a certain job may not want to answer a specific question) or calculating an appropriate replacement value based off other observations.

3.1.2.4 Data analysis

The analysis aims to distinguish if an occurring event reflects a true effect or a false one (RCR,2020).

Once the data is thoroughly cleaned, analysis can begin. First data is manipulated in base ways to try to find correlations (Simran,2020).

Once data manipulation has been performed then it is possible to get sufficient information to answer the question however, it's plausible that the results gotten may not help in answering the question, this means the question needs to be rearranged or more data should be collected. This process is repeated until enough information has been retrieved to answer the question. Analysis should be accurate and appropriate to ensure the integrity of the data. Poor statistical analyses would result in distorting of findings which could have negative effects. Susan Etlinger stated in her 2014 TED talk (Etlinger, 2014) "At this point in our history... we can process Exabyte's of data at lightning speed, which also means we have the potential to make bad decisions far more quickly, efficiently, and with far greater impact than we did in the past."

3.1.2.5 Data interpretation

The information received from analysis must be interpreted. Results retrieved should be compared with the initial question to see if it provides an answer. Data interpretation also gives you a chance to reflect on the analysis that has been undertaken, check that all possible parameters have been considered or if there has been any hindering factor that possible has been implemented that affects the overall results. Ethically, it is important that the interpretation is unbiased and not trying to cause deceit. Doing so would mislead the casual readers (Harriss, 2018), and may negatively influence the public perception of research.

3.1.2.6 Data visualization

It is very common that after analysis has been done that the results get presented in a visual medium as it allows for a better understanding of results. Typically, the data is transformed into a chart or graph. Data visualization can also use to discover meaningful information that was previously unknown (Dillard, 2020).

3.1.3 Reliability of analysis

Gottschalk (1995) identifies three factors that can affect the reliability of the analyzed data. When any of these three factors cannot be consistently demonstrated it suggests that there are integrity issues.

These are

- Stability – A data set is that is analyzed is a certain way should consistently return the same results.
- Accuracy – The analysis of the data should be able to produce outcomes that accurately portray the reality of what is happening.
- Reproducibility – a way that the data has been analyzed should be able to be repeated by another researcher.

3.1.4 Data analysis and optimization

There are many cases where data analysis has been applied to a problem to optimize a solution. There are even companies are to be hired to do data analysis for other companies. One such company is edge analytics. Edge (Edge, 2020)

One such case is when they were asked by a government emergency service to analyze how they could optimize the placement of their response hub to minimize time and travel. By using the average travel time between each point edge analytics developed a bunch of scenarios of where they could be placed.

3.2 Energy consumption

As my dissertation is going to involves creating an energy policy, I decided it was important to research current energy policies.

Energy consumption refers to the amount of energy used to perform an action. In this section, I will discuss energy wastage, what an energy policy is and current policies that have been created in order and save energy.

3.2.1 Energy wastage

Energy Waste is the consumption of electricity by a device that is not performing any useful action. Energy wastage usually occurs when a device is poorly maintained or when the device is left in an idle state for a long period whilst not being used, for example putting a TV in stand-by-mode just before you leave the house for a few hours. Assuming that you are the only one who can access the TV, putting the TV in a state where it is ready to be used when you know that it is not going to be used for a few hours is just going to waste energy. Putting it into a sleep state would have been more beneficial.

Energy wastage can be split into three different areas

3.2.1.1 Long Term

These are permeant problems with a device that leads to the device leaking energy. For example, maybe an issue has caused a PC to have its fan running at all times. Long-term problems can be the result of poor device maintenance, this is an issue as most electronic devices consuming 20% more energy when it is in poor maintenance (venkatesh, 2013).

3.2.1.2 Regular

These predictable daily events cause the device to waste some energy. I.e. A TV is left on for a few minutes while you go and do something else.

3.2.1.3 Intermittent

These events happen unexpectedly and cause a device to waste some energy. The TV accidentally left on overnight.

3.2.2 Energy conservation

Energy conservation means reducing the amount of energy used by a device in a smart way, by either using incoming energy in a more efficient way or by putting rules called polices onto the devices.

3.2.3 Energy policies

“Energy efficiency policies and programs can help drive the implementation of projects that minimize or reduce energy use during the operation of a system or machine and/or production of good or service.” (Energy.gov, 2020)

For appliances, there are minimum energy efficiency standards enforced by federal governments. These standards enforce that the products must adhere to a certain maximum energy consumption rate.

Energy policy in most devices the time that the device must wait before entering a sleep mode is either set by the owner of the printer or by default predefined by the device manufacturer according to Energy Star (Energy star , 1992) environmental standards. Energy Star is a program run by the U.S. Environmental Protection Agency and U.S. Department of Energy that promotes energy efficiency. Most devices have been made to meet their standards.

3.2.4 Device modes

Devices are modeled using the following modes.

3.2.4.1 Active mode (or busy mode)

The device is currently doing one of its active functions. The device will stay in this mode until it’s told to stop (i.e turning a TV off) or until the device has done its job (i.e automatic alarm clock). In a printer this when a print job has been passed, the printer activates its marking engine, print path and controller and then completes the job. Typically, power consumption is the highest in this mode.

3.2.4.2 Idle mode

The device is waiting for it to receive some input in order to activate itself immediately. In a printer the idle mode has the device waiting for a new print jobs, once it receives a print job it will automatically print it off. This means the device is on but not in active use. A certain amount of power is required to maintain the printer in a state of readiness. The time a device waits in idle is called “time before shutdown”

3.2.4.3 Sleep (or power-save) modes

The device is still on, but not readily available meaning time and energy will be needed to get the device to the active mode. When a print job gets passed, if a user prints to a printer while it’s in sleep mode, then there is a slight delay before the document is printed.

Typically, the device will only have one active mode and will have one or more sleep modes.

3.2.5 States transition

There are names for the transition between states.

3.2.5.1 Done

The transition between active and idle

3.2.5.2 Activate

The transition between idle and active.

3.2.5.3 Wake up

The transition between the current sleep to active mode is called the wake-up.

3.2.5.4 Shut down

The transition between idle to sleep is called shutdown.

3.2.6 Time related phrases

3.2.6.1 Time-out

The amount of time a device should stay in idle mode

3.2.6.2 Break-even

The amount of time where a device must stay in the sleeping state long enough to compensate for the wakeup energy overhead. These can be called the minimum duration and the minimum sleeping time. The minimum length of an idle period to save energy is the break-even time for an idle period.

3.2.6.3 Threshold time.

The point in time in which switching to the next mode is more efficient than remaining in the current mode. In this project printers with one sleep mode are being tested so the threshold time will be the point in time which more energy has been used staying idle than the energy used by the wakeup process. It is more efficient to be in sleep mode if we know the next job is greater than this point in time.

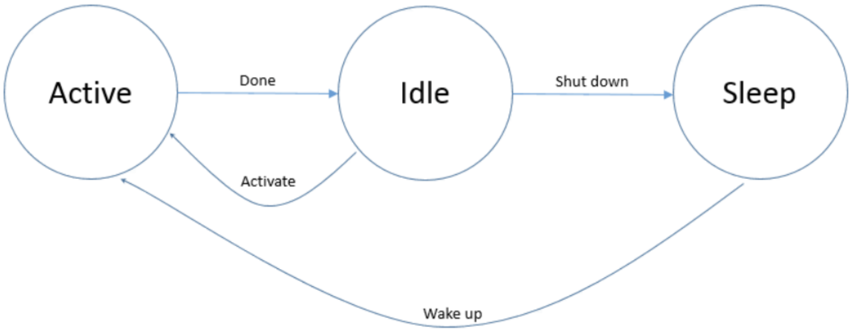


Figure 2 – A diagram represented of the modes and transitions of a single sleep state device

3.2.5 Consumption model

When we want to work out the energy consumption of a device then we can model it the follow way.

For device with multiple sleep mode we use the following annotations

The number of sleep modes	M
Idle mode	I
Current Sleep mode	J = N
Power consumption in idle mode (watts)	B
Power consumption in sleep mode J (watts)	C_J
Energy required to switch from Sleep mode J-1 to J (Joules)	D_J
Energy to wake up from Sleep Mode J (joules)	E_J
Power Consumption in Active Mode	A
Time between last job and now	T
Time between current job and next job	S

It assumed the sequence J0 , J1 , ….. , M is always going to be decreasing in energy consumption and that the only state a sleep mode can transfer too is J+1 or 0. As we know our printers to be of single state I will be modelling all polices based around that fact.

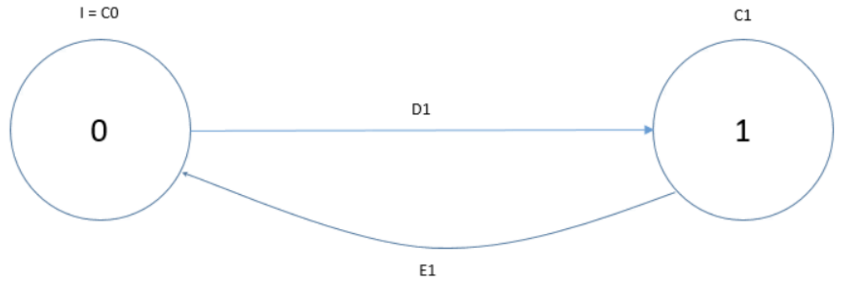


Figure 3 – Consumption model of a single sleep mode device

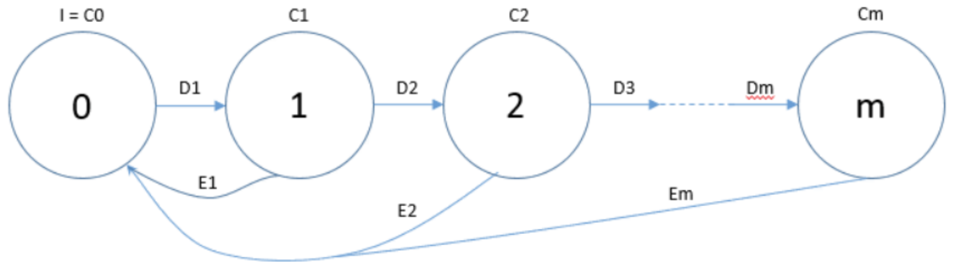


Figure 4 – Consumption model of a multi sleep mode device

3.2.5.1 Order of power consumption

$$A > Ej > B > Dj > Cj$$

Power consumption is highest in the active mode and the lowest in sleep mode. The transition from idle to active takes less energy than the wake-up transition. For devices with multiple sleep modes then the required power used in wake-up increases the deeper, the device goes through its sleep mode.

However, jumping from Sleep mode to active typically takes more time and often takes more power than switching from idle to active. If a device is always finding its self-needing to be activated just after the idle timeout has passed, then more energy will be used than just waiting in idle mode for a longer period.

3.3 Current energy saving algorithms

With the consumption model, set up I will now talk about current energy policies that have been created to reduce energy waste by setting an optimal idle time for the device.

3.3.1 Timeout

A simple and well-used algorithm that a lot of devices use and seems to be the standard due it being used in most energy start certified devices (energy star, 1992). The logic behind it is if a device has been idle mode for a long period then it's probably going to remain in idle mode for a long time. Timeout works by setting a point in time that when current time T exceeds this point it will switch to sleep mode.

$$Mode = \begin{cases} Sleep\,, & if\ T > Timeout \\ \ Idle\,, & otherwise \end{cases}$$

3.3.2 Exponential average

The exponential-average approach (Peterson and Silberschatz ,1986) is one of the commonly used methods for the prediction of the idle period.

By using the pervious predicated and actual lengths of the previous idle periods, create a timeout predication for the next idle periods. The prediction is made by using an average of past idle time with added exponential weights.

The following formula is an modified exponential-average approach that us used in the CPU scheduling problem used by Hwang and Wu (Hwang and W, 1997) in when they looked into using it to create a more efficient energy policy.

$$I(n+1) = \big(a * i(n)\big) + ((1-a) * I(n))$$

I(n+1) is the new predicated idle period, I(n) is the last predicted idle period and i(n) is the latest idle period, a is a constant attenuation factor in the range between 0 to 1 that affects the relative weight of the past or current history jobs. If a = 0 then-recent history has no effect and the policy will be based purely on a prediction. If a = 1 then the only recent history has an impact and the prediction will be based only on history.

3.3.3 L-Shape

If putting all of the times between jobs on scatter graph an "L-shape" pattern will occur made up of short busy periods followed by a long idle period. If this L-shape is observed, then the predicted idle time should match the L shape pattern (Srivastava,1996).

The problem with this is a sudden long idle period after continuous, near-uniform idle periods would throw a predicted idle period that's much lower than the actual idle period (In > In`) meaning the device will waste energy in an unnecessary idle mode. However, on the flipside the predicted value for the short idle period after the long idle period could be correct due to overshooting (In < In`). A way to solve this using the formula developed for exponential average is to include a monitor that would periodically record the current idle time. It would predict the idle time and If the predicted value is less than the threshold time the device should wait in idle mode up until the threshold. Once this threshold is reached and there hasn't been a print job a second prediction shall be made. If this prediction is greater than the breakeven threshold then the device should go to sleep, else it should remain in idle mode.

For example, we have a long idle period I5, if I5` is less than the threshold time then the system stays in idle. After the threshold another prediction is made, if I5`` is greater than the threshold time then the device goes into sleep mode.

$$Mode = \begin{cases} Sleep\ , & if\ Threshold\ <\ \left(\left(a * i(n)\right) + \left((1-a) * I(n)\right)\right) \\ Idle\ , & Otherwise \end{cases}$$

3.3.4 Competitive algorithm

The competitive algorithm (often also referred to as C-competitive power saving algorithm) is applied to a system with several states. It is based on setting the timeout to the point in time where the cost of remaining in one state out ways the cost of switching to another state. However, in this project the devices used only have 1 sleep mode. Competitive algorithm works for devices that only have 1 sleep mode however as seen in A. Karlin (June 1994) this seem to use twice as much power when compared to other algorithms.

$$Mode = \begin{cases} Sleep, & if\ (T *\ Cj) > Ej \\ Idle\ , & Otherwise \end{cases}$$

3.3.5 Oracle average

Oracle average is an algorithm in which the future of the device is assumed to be known. In this project, the printer would switch into sleep mode immediately if the between the current and next job is greater than the threshold time (it is cheaper to stay in sleep mode and then wake up than stay in idle mode).

$$Mode = \begin{cases} Sleep, & If\ S\ > \text{Threshold time} \\ Idle, & Otherwise \end{cases}$$

3.3.6 Exhaustive search

The Exhaustive search algorithm is a method that consists of finding the time out that minimizes the actual consumption of energy based on the data set at hand (Weinstein ,2020). To get the most optimal constant timeout a grid of possible values is considered. The actual consumption is computed for each possible time out, with the optimal one being retained.

3.3.7 Learning tree

By modelling discrete idle period as leave nodes on a tree, a finite state machine can be used to predict the following idle periods and selecting the branch that best matched the situation. This algorithm can be used on devices with multiple sleep states as at the beginning of an idle period the algorithm can select the appropriate sleep state (Chung, 1999).

$$Timeout = \ (Predicted\ time\ diffrence)$$

3.3.8 Stochastic model

By modeling devices as stochastic processes, power management can be solved as an optimization problem, meaning it can provide flexibility when wanting to tradeoff between power and performance. This is done by looking at the optimal probability of each action, by looking at past actions and the amount of power it consumes. This can be further improved by modeling device as a continuous-time Markov process. This allows the changes to happen based on events and not times, allowing for decisions to be made as soon as possible. However this has led to problems as mentioned in quantitative comparison of power Management Algorithms (Hsiang , 2008) as the algorithm tends to shut down too soon, thus incurring large wakeup costs and sometimes even missing shutting down on long idle periods” (Simuni ,2000). Variants on this process are the Semi-Markov approach (which relaxes some decisions by allowing for state probability deterioration) and the piecewise homogeneous Markov processes (the begin of the initial state sets up the time for the shutdown, during the process it evolves in a Markovian way. During shut down the initial state of the following process is created). T

3.3.7 Interesting finding

There was a paper written by Hwang, Chi-Hong and Allen C.-H. Wu that used a range of the policies discussed here, timeout, l-shape experiential average and the Threshold. It came to an interesting conclusion that if the if the energy consumption needed to change to device is high then standard policies of long timeouts are probably better. (Hwang, 1997).

Chapter 4 Data preparation

In this chapter, I will discuss the data received and what I did to make sure that the data was clean.

4.1 The data

The data is a record of user logins and another data set of print jobs. The login data contained logins for the year 2019 whilst the print job for 3 cluster rooms over the years 2014 - 2019.

4.1.1 The Cluster rooms

L1 Tees

This is a small out-of-the-way cluster room has that has 19pcs and one black and white printer called 133.

L1 Hope

This is like the Tees room; a small cluster room that has 24pcs and one black and white printer. This room can also be booked out as a teaching space or other similar reasons called 123.

L4

A large open floor cluster. There are 80pcs however there is also space in this room for users to set up their own devices. This room has a separate B&W printer and a colour printer. The B&W printer is called 95, the colour printer is 96.

These are not all the clusters rooms that are in the building. As far as I know, all cluster rooms follow the building opening hours and closing hours and all rooms are open and closed at the same time (i.e. holidays).

The print job data starts at the 2014 academic year (September) and ends at the end of the 2019 calendar year. For the most part, I will be excluding data from 2014 as it only contains a few months of print jobs so when looking at months it would have skewed the results a bit. When I am looking at academic years and semesters, I will include the part of the 2014 year that's needed to create the full 2014/15 academic year and remove the end of 2019 (past September).

The login session datasets contained all logins in 2019 for all cluster in the building. As I only had print data for the 3 cluster rooms, I filtered the login data to merge it to the print data to create another data frame which was all the print jobs matched to appropriate user sessions.

4.2 Clean up

Before I did any analysis with this data, I followed the standards set out for doing data analysis and made sure I took time to clean it to make it technically correct and then further cleaned it up so that it was consistent data.

When I was making it technically correct, I made sure that my columns were suitably named, and that each column had an appropriate R type that matched the value domain of the column. For the Print job data, I added in an additional column where I had matched each job to the correct cluster room to add more data that I could analyses. I also made sure that all the values in each column shared the same unit of measurement. I broke down the timestamp into the separate time functions (Year, month, day, hour, and minutes) to allow for easier analysis. After making the columns have suitable R types I then checked the dataset for any NA values and found none.

The next step was to remove fully duplicated rows. In the print job dataset there where around 3,000 rows that had been duplicated (out of 600,000 + records). From scanning through the duplicates there didn't seem to be any correlation between the duplicates, only that there were a few days in which there was 5+ duplicated print job. There was one day in June 2019 that contained so many duplicate jobs (2283 duplicates) that it singles handily brought up the month's average up by a significant amount. I solved this issue by removing these duplicated rows from the data set.

The only column that I cared about uniqueness was the timestamp of the print job printer. A printer shouldn't be able to print off two different users' documents at the same time. After looking to see if any jobs matched this, I discovered there were 2,000 cases of two separate jobs being printed off at the same time. After some investigation of these duplicates, I decided to keep them in. I decided to do this as printers are made to handle multiple jobs via the use of queues. In reality there would be a few second delay between these jobs. To fix this properly I add a slight delay of a few seconds to the second job.

I removed some rows from my initial dataset that I felt would be irrelevant to my overall data analysis, for example, there was a set of two printers that used to exist in the L4 cluster room that where removed in the year 2014. All the other printers existed across the whole data set meaning it comparison using these two would have been pointless.

I also discovered that the colour printer only handled colour prints suggesting there is some background force that stops the colour from taking B&W jobs.

When searching for the number of slides printed off by a user, I discovered that there were about 5000 cases of zero number of sides being printed out printed. This suggested that the printer processed a print job but didn't do anything. Originally, I thought this might be due to people scanning a document out however I discovered that none of these devices has that function. I'm honestly not sure what caused zero number of page prints however the printer did see this as a job, thus it would cause the device to become active. Because of this, I kept these values in as it still a case of the printer waking up.

The print jobs are one-time events and the whole dataset essentially a log file of print jobs stored by the printers or some external system. Because of this, it is difficult to check that the data is 100% accurate as it would require another source of information to back up the time of the print jobs like interviewing the person who printed who did the print job. However, the data is anonymized and even if I were to get in contact with the user, they may not remember the specific dates and time when they printed something off.

When I got the login data, I knew that at some point I would be comparing it with the print job data. Because of this, I spent a fair bit of time getting the column in login dataset to match the name and type of the print job dataset to allow for easier comparison. Another way I cleaned the login data was to remove many irrelevant data. Login data contained data for all clusters where I only had the print data for three cluster rooms, so I removed the login jobs that I wouldn't need. The logins didn't originally have the cluster rooms that each PC was in so that was something that I added.

A big issue that I came across was print jobs data set with the login data set. The first issue was that the logins dates set didn't match the British summertime meaning that when the clock went forward all the logins were an hour out meaning that out of 80,000 print job that happened in 2019 more of half them didn't fall within a user's login session. After editing the data to better represent BST this number went down, however, didn't reach zero. After further inspection there were 9,000 jobs were a user print job has gone through however the user has already logged out. Most of these jobs are within 30 minutes of the user login off which could be a data issue, or it could be an issue with the printer. From personal experience, the devices can sometimes have a delay when they are printing something out, also you tend to see many documents that have been printed out but never picked up, and this also suggests delay. There was also the issue of around 20,000 print jobs not having a matching login in the login database. People printing using their own devices, which is possible due to a university website that allows a user to print to any printer from their personal device, could cause this. A student ID needed to print accesses this site so the print job would still associate with a user.

After deciding to look at the time between jobs to be important I noticed that there some time differences that were outliers when compared to others. I kept these in as I felt that they were needed, as at the end of the day they are real-time differences between two print jobs. Removing them would mean I would be ignoring them completely and that may be more damaging than good.

There is another issue of a week worth's of print records are missing mid-2019. Though I could generate a replacement set of print jobs my prediction could have been wrong. I decided to leave this week blank, as it was the safer option.

Once I felt that I got all the issues with data sorted, I was ready to start the analysis.

Chapter 5 Analysis

In the following chapter, I will discuss my analysis and how it helped me develop my policies. I look at the differences between the current print job and the next print job. What I am trying to optimize is very time-sensitive, if a printer goes to sleep just before a new print job is passed then more energy is going to be used. If we can work out the times between these jobs, then we could potentially solve this issue.

5.1 Type of analysis chosen

I decided that for my analysis I would be primarily using statistical analysis. Time can be represented in two ways, discreetly as a count (I.e. number of jobs per hour) and continuously as the time between time jobs, so using a technique that would allow me to take advantage of that fact was important. As we are trying to find the key timings, Statistical analysis will allow me to find means and other quartiles that could be helpful in solving the problem.

5.2 Time

In this chapter, a brief analysis of the data will be undertaken to develop a greater understanding of the printers and when they are typically used. As policies revolve around a timeout period, I wanted to isolate the factors that would cause a difference in the median time out between print jobs. I used the Median as the determining factor over average as I quickly realized that there was a handful of outliers in the times between print jobs which would messed up the average. These were primarily the print job at the end of a user’s login session who was also the last person to use the printer for a while.

5.2.1 Month

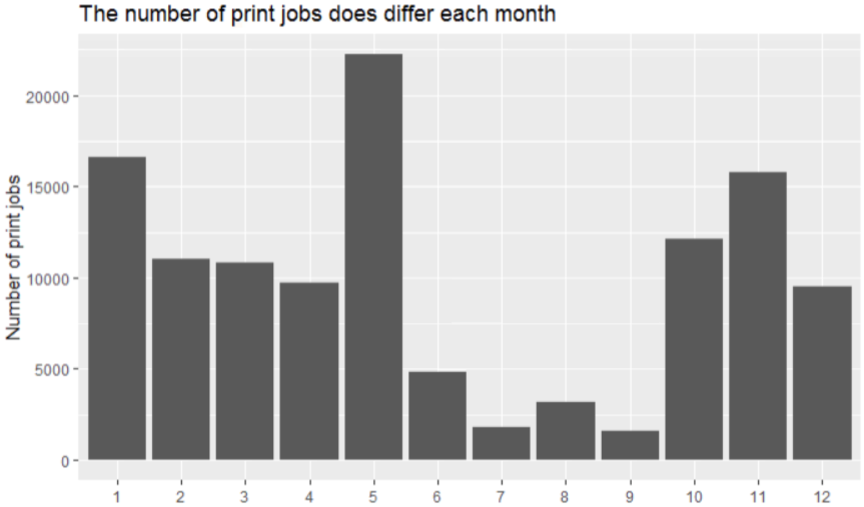


Figure 5: Bar charts of the average number of prints jobs done each month

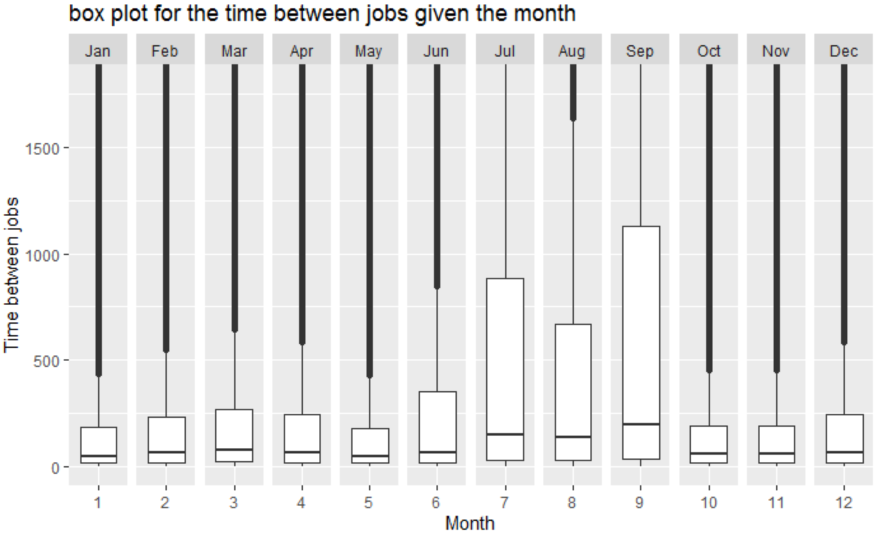


Figure 6: Boxplot for average number of prints per month

Months seemed like the logical first step. By getting the total number of jobs per month I created two graphs, the first being the average count of jobs done in each month and the other showing a box plot of the time between each job given the month.

I was expecting there to some variance between the months however, I didn't expect such a wide range of difference. The median boxplot graphs seem to produce the inverse of the bar chart.

The first graph shows that the month defiantly influences the total number of print jobs. I believe that this may be due to events that occur during the academic year. I discovered that there is a fair a bit to talk about so further semester analysis will be discussed in section 5.2.

The boxplots also tell us a lot. There is little difference between the medians for each month. The lowest median was in January at 50 seconds and the highest in September at 194 seconds (a 2-minute difference).

This shows us that 50% of all jobs for each month where printed within minutes of one another. However, the boxplot generally differs with their maximums, with some month's maximums being less than other months' upper quartiles. This shows that while the 1st 50% of each month prints share the same values, the other 50% may vary.

Examining both graphs at the same time shows that months that had more print jobs had, for the most part, less time between print job. The months where this can be clearly seen in the summer months, which is usually a time for breaks, suggesting that there will be noticeable change between the standard work times vs break times.

When I think of when a user is most likely to take a break I think of the weekends and nights (after uni ends for the day for most people). This theory is why I am going to investigate differences between weekdays and hours next.

5.2.2 Weekday

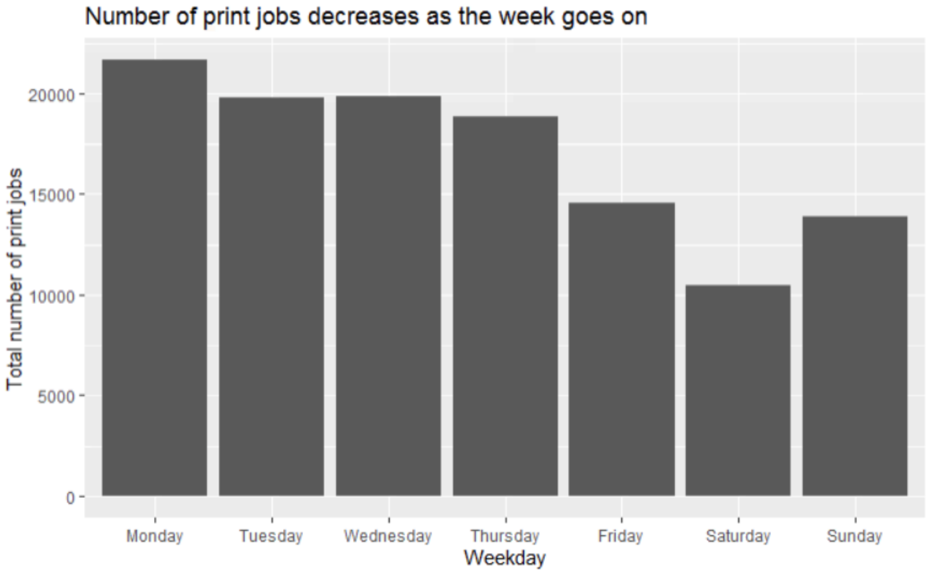


Figure 7: The number of print jobs done each day of the week

Next, I wanted to look at the total number of print jobs done on each day of the week. The bar cart is meant to represent the total number of print jobs in each day for average year. I expected that the weekends would have noticeably fewer print jobs than the weekdays.

The trend that the graph shows happens in all years, more printing at the start of the week. This could be because of people printing out documents to that they will need later in the week. There seems to be a big difference between Thursday than Friday then Saturday with a noticeable increase on Sunday. Again, going off my previous possible explanation then it is likely that people have already printed off the document that they need for the week, so they have less need to use a printer.

The increase on Sunday is probably due to people wanting to print documents off to prepare for the following week.

People slowing down for the weekend may cause the dip towards the weekend. There have been many studies on the correlation between weekdays and productivity and each show the same results, the further we get into the week the less productive we get. This matches the results the graph shows.

5.2.3 Hour

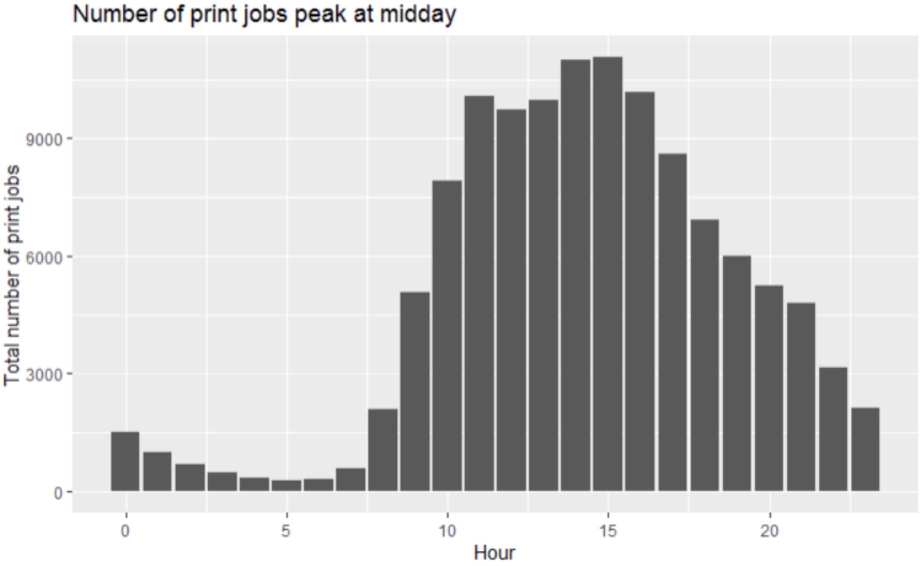


Figure 8: On average over a year many print jobs where done in each hour

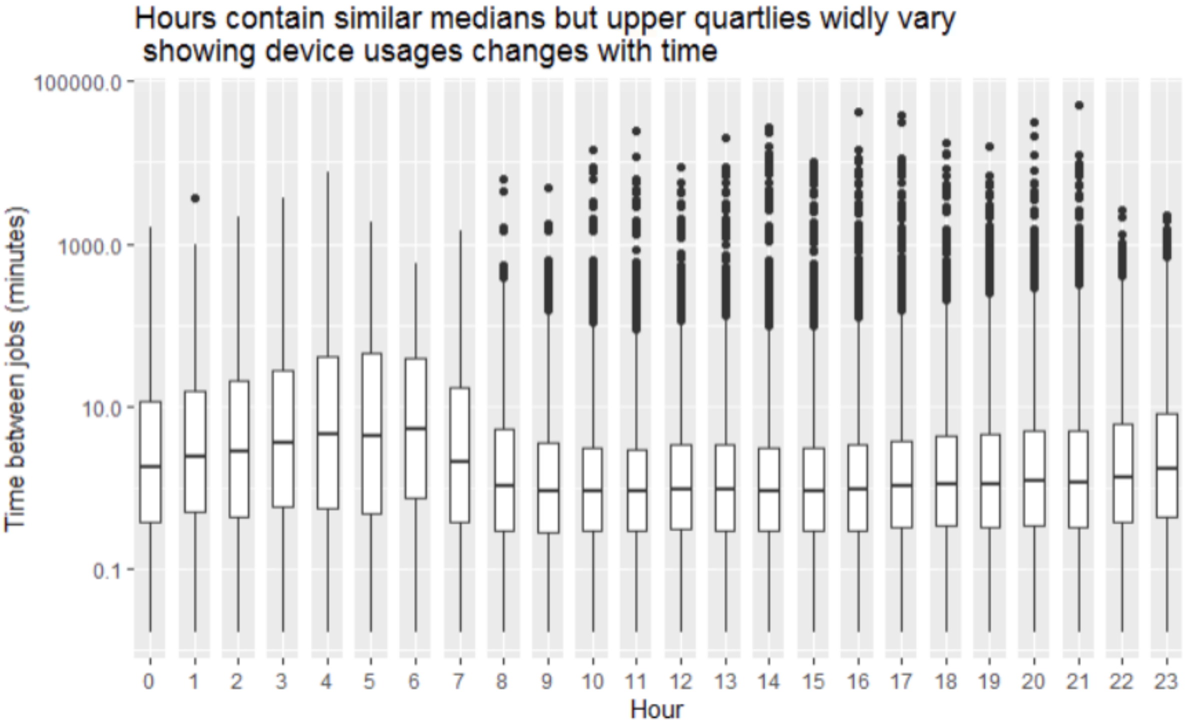


Figure 9: Hourly box plots

After looking at weekday, I decide to break it further down into hours. Like the previous graphs, the times that are normally thought of as off-hour that has fewer jobs done whereas active hours have more. There is a noticeable rise at 8 o'clock, this is probably because this is when staff tends to come and an hour before lectures tend to start (which start at around 9). The number of jobs per hour increases and peaks between hours 13-15. However, it has should mention that there is seems to be a drop in the total number of print jobs around the time people would go for lunch (11 – 13). The number of prints jobs per hour seems to steadily decrease after 16. This is the last working hour for most university staff and the around the time of the last lectures of the day. The box plot shows the break hours tend to have longer times between jobs than the active hours.

5.2.5 Timings combining

As we have seen in the previous section that time does play a factor in device usage, whether it be in the form of the hour, weekday and month. There is busy times and slow times. All areas of analysis in this section were then combined to get the median for each hour in a weekday for each month. This information would be used for my first policy as it's safe to reason making a policy that is based around the expected time between device usage would be a good start.

As timeouts is a set amount of time in minutes, I investigated the effect minutes have but found it did not produce any useful information and there wasn’t much to talk about.

5.3 Semesters and Holidays

In the last chapter, it became clear that device usage changes depending on the month and I theorized that this might be due to the events that happen over the academic year. In this chapter, I will discuss my investigation into semesters, breaks and holidays.

Though the dates change each year, the academic year can be split into 6 sections

Semester One: late September to mid-December, then early January to late January

Winter break: mid-December to early January

Semester Two: Late January to late March (though this has fluctuated in some years to early March)

Easter break: Early March to Early April or Late March to Late April

Semester Three: Early April or Late April to Mid-June

Summer holidays: Mid-June to late September

Two big events take place during the academic year, the exam periods

Exam period 1: Mid-January to late January

Exam period 2: Mid-May to early June

I wanted to find out which semester was the most important; the following table is the average percentage of the academic year each term takes up and the Percentage number of total jobs done in each semester.

Semester Time	Percentage Year	Percentage total	How much of the total print jobs dose each percent of the semester makes up?
Semester One	26.9%	42.64%	1.5
Winter Break	6.1%	3.4%	0.5
Semester Two	22.6%	21.13%	0.9
Easter	7.9%	15.32%	1.9
Semester Three	7.3%	11.43%	1.5
Summer	27.8%	5.59%	0.2

When looking at how important each semester is, I made a weighting based off the Percentage total divided by Percentage year. Though it has one of the smallest percentages of the year, Easter has the highest ratings by a decent margin. Tying for 2nd is Semester one and three with a 1.5 weighting, this makes sense as a lot of deadlines are primarily in semester one and three, also both contain one of the exam periods. What should also be mentioned is that both semesters make up wildly different percentages of the academic year. This tells us that the length of the semester/holiday doesn't correlate to device usage, instead, it proves that it’s more tied to key events that occur in the semester.

To further this idea, I decided to do a day by daybreak down of the 2019 academic year. The following a graph off all the print jobs done by all devices.

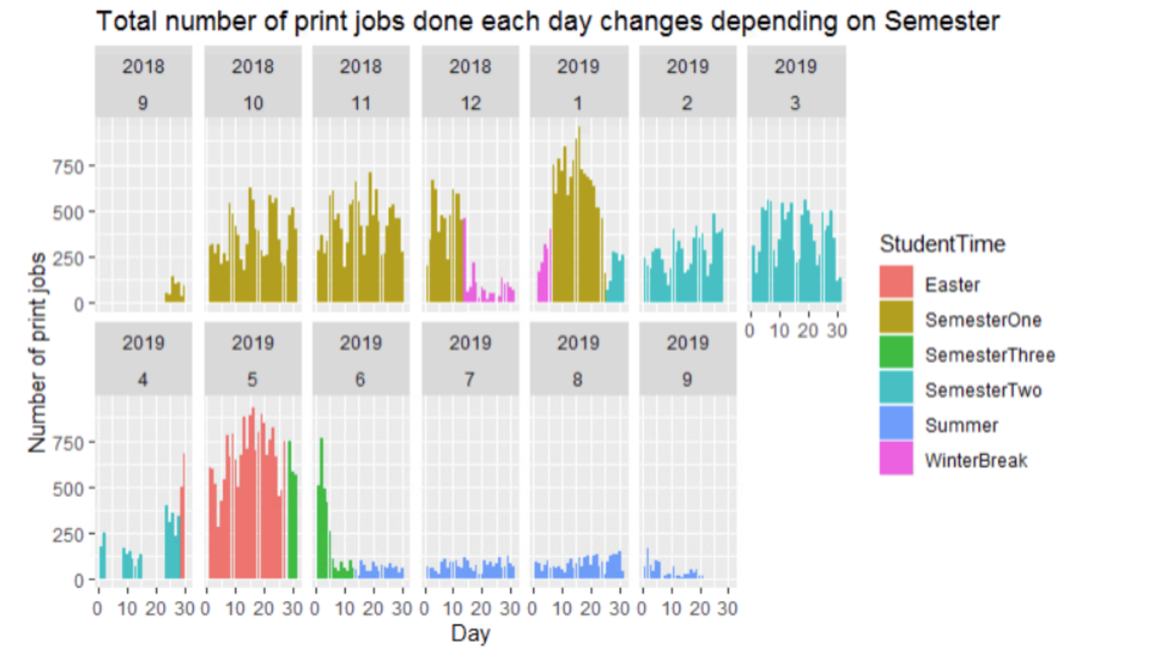


Figure 10:Number of prints done each day semester

Summer and Winter break have a low print count. Summer makes sense to have a low print count as that is after semester 3 and is when the academic year ends for undergrads. Most undergrads won't have any coursework or exams to do and normally won't need to be on campus. The lull in winter may be for a similar reason. It's the winter holidays so most student will have left to go home/have a break so again probably wouldn't on campus. Easter is also a holiday in the academic year but makes up a relatively large percentage of the years print percentage, this may be caused by it being right before exam period or it might because a lot of courses set reports to be completed over the Easter break. If those courses have paper hand-ins then it's likely people will be printing off their coursework over this period.

There are spikes in the printer usage within each semester. After further investigation, this lines up with the weekdays, with most the spikes being Mondays then it slowly drops throughout the week, with a respite at the weekend. This is the case in most of the time, however, this weekly print trend changes during the exam period. This can clearly be seen mid-January. It seems like at the start of each of the exam periods the number of print jobs increases no matter what day. There is also another interesting trend where there is a certain peak in exam period where the number of print jobs starts to slowly decrease until the end of the exam period, then it rapidly drops. This can most likely be equated to the fact in most situations that once a person has done their exams then they are done with the module and won't have any course work till the next semester, meaning that they most likely are taken a break. This can be seen after the second exam period. As once the exams are done there a few print jobs.

Another clear observation is the changing between semesters can cause a large increase or decrease in the number of print jobs. Take the change between semester one into the winter break back into semester one. The isn't a gradual transition, as soon as it becomes winter the number of jobs drop over winter and then when the semester starts again the number of jobs jumps back up. A similar trend can be seen between the end of semester one into two and the end of semester two into semester Easter. The only case where this does not seem to exist is between the Easter holidays and semester 3.

After looking at each the transition between semesters every year, they all followed the same trend of peaks and respites suggesting that these are recurring events that that will occur in future years.

This shows setting up a policy based on months would lead to slight inefficiencies as one month could contain so much change due to semesters switching that having flat rate would use the middle ground between two very different events. This suggests that having a policy changing based on the semester (with special cases like exam dates periods takin into account) would be beneficial.

5.3 Devices

The data collected contained information for four printers that are located throughout three different cluster rooms. As we are making policy to optimise a device then it would be beneficial to analyse the individual usage of each device. Due to the difference in size and styles of the room, I expected that there would be a difference in printer usage.

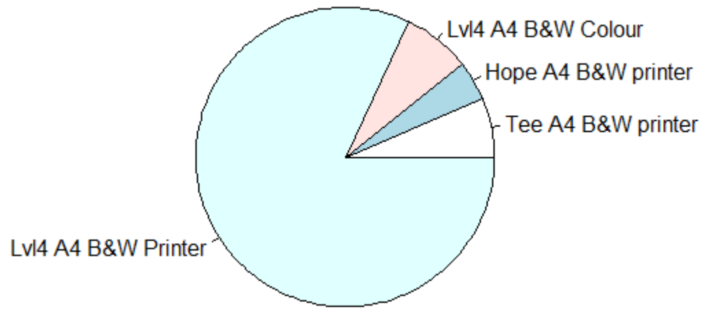


Figure 11: Percentage contribution that each printer has made to the total number of print jobs.

When looking at the percentage number of jobs done by each printer, it is clear to see that one printer makes up most of all print jobs, with the others all have similar percentages.

The cluster room influences the device as out of the B&W printers the one in the biggest cluster room (Lvl4) had the greatest usage with the other rooms of similar size having similar percentages. However, there is a clear difference between the usages of both devices in the LV4 cluster room with an 11:1 ratio. This shows usage depends more on the device and not the cluster it is located in so a policy tailored to each device may be the way forward.

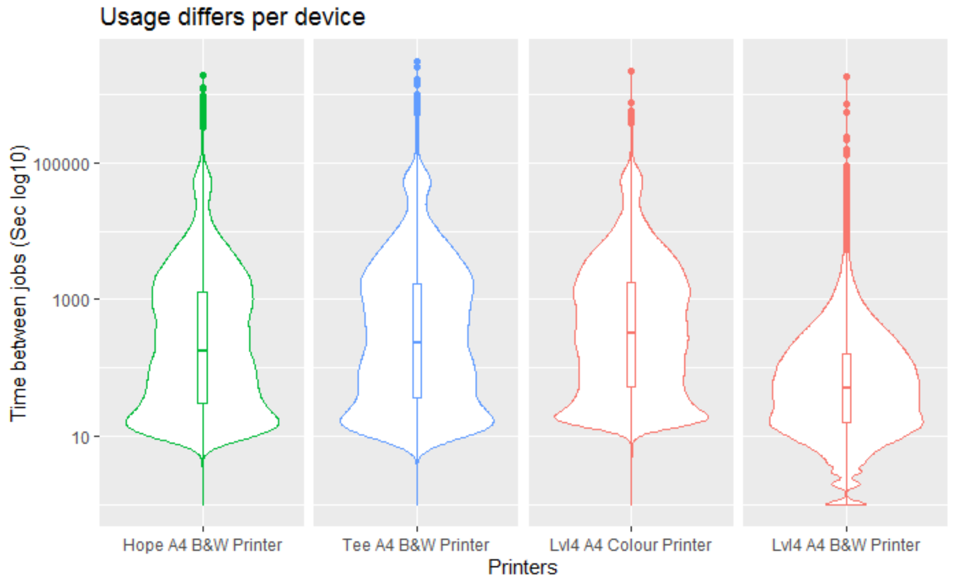


Figure 12:A violin plot of the time between each job - Printer

When converting the time differences between jobs into violin plots we find that those printers which shared a similar percentage of print jobs show similar violin patterns, a wide bottom slowly curving into a peak. The two devices in the same location have different device usage and produce very different violin plots. The medians are very different as well, with 75% of all B&W printer jobs taking place before the median of the colour printer.

When comparing the time between print jobs its clear to see that policies should focus on the device

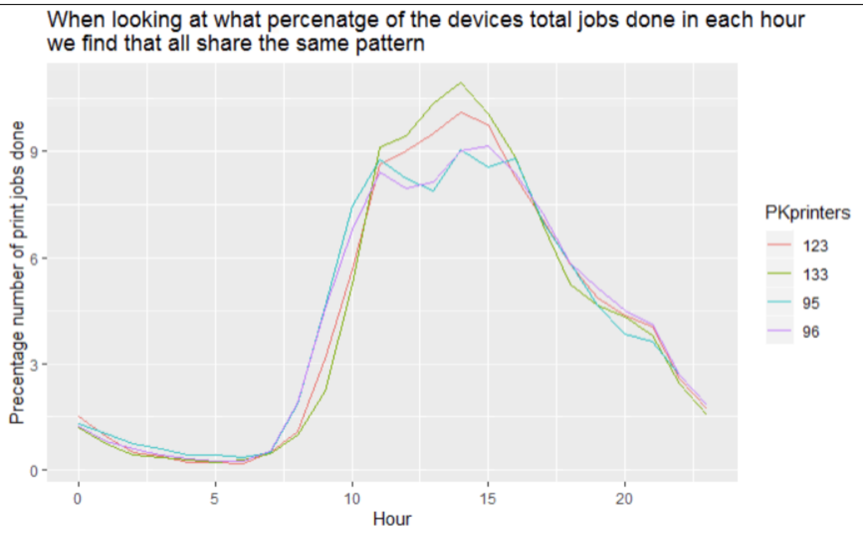


Figure 13: Percentage of total number of print job for a printer by hour

When we look at the percentage of devices total print jobs done by each hour, we find that they all follow the same pattern. Suggesting that all the devices will still be used in the same ways at the same time but just in different amounts, meaning previous finding will also apply and could benefit each device.

5.4 A user's logins and prints

Originally, I was going to be able to simulate when a user logged on however due to changes I won't be able to. I did a fair bit analysis into this but in the end; some of it had to be scrapped. I have included it in appendix 2. My main finding from this is that most users that print something off are logged on for less than 15 minutes suggesting that when someone logs on then a print job going to happen soon.

5.5 Multiple print jobs

When looking into how many people print off, I discovered that over the 5 years that there were 91,512 instances (41%) of a user only printing off one document whereas there were 129,786 instances (59%) of people printing out 2+ documents.

Though there are clearly many 2+ jobs, which suggest that when a person logs in to one of the devices they are likely going to print multiple documents. However, there is still a significant number of people who only print off one document. Whilst it may be easy to work out when a user is going to print off another document if they have already printed out a bunch, we still need to be able to create a policy that deals with people who are printing off just one document.

Min	1st Qu	Median	Mean	3rd Qu	Max
1	19	64	1556	375	84236

When getting all user session that contained multiple print jobs, I worked out the median score was 64 seconds between each print job. This suggest that when a user is going to print off multiple documents then they will have a set of documents prepared to print all at once.

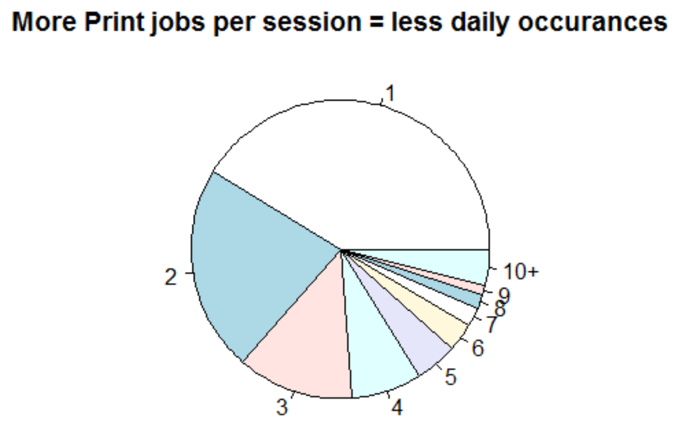


Figure 14: What percentage of did user printing x number of documents make up.

That then lead me down the path of would the more a user prints out effect the time out between jobs. As you would expect the higher the number of print jobs N a user does in one user session the lower the daily occurrences of a user printing off N document in one logon session. As the number of prints over 10 gave the same percent, I decided to merge them together. Further investigation backs this up as jobs over 10+ shared a very similar time between jobs. Doing it this way made it easier to analysis and turn into a policy.

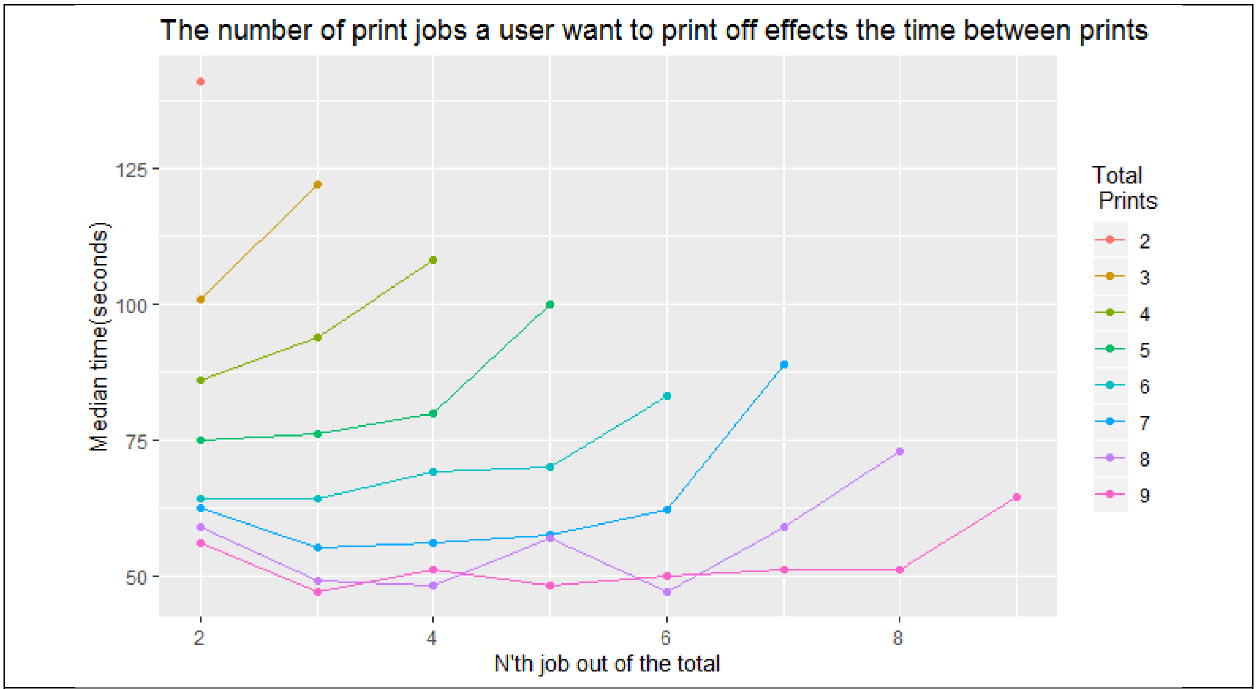


Figure 15: Given the amount a user prints, what is the median between each job

This graph shows the amount the median time between a users print job based on the number of jobs that they have printed out. It shows no matter how many documents a user prints off it’s probably going to be done in a burst. So we can say if a user has printed some documents in quick succession its likely if there next print job will be done sooner than later.

Chapter 6 Testing & Results

The main aim of this dissertation is to create an optimal energy consumption policy that balances energy-performance trade-off by using information gathered from analysis to develop a policy. Policies created should cause the device to use a low amount of energy whilst also sporting a low inefficient device turnoff.

6.1 The Simulation

Using the print data of all jobs done in 2019, I ran several different policies through my Java-based simulation to discover how much energy was used between each print job.

The printers where modeled off their real-world energy usage.

Mode	M608	M651
Printing (Watts Hour)	780	650
Ready state (Watts Hour)	45	67
Sleeping state (Watts Hour)	0.8	8.5
KWh / week	2.622	4.7
Joules need to go from sleep to active based off Phaser 4500 model	21.3kj (6 w)	21.3kj (6 w)
Joules need to go from idle to active based off Phaser 4500 model	1 w	1 w

Due to unfortunate circumstances, I am unable to get the amount of energy in joules used to transition between sleeping states to active for these printer. However, I was able to find the amount needed for a similar single sleep state printer called the Phaser 4500. I decided use a slightly reduce number because the Phaser was brought out in 2004 so it is probable that advancements in technology would reduce the amount used by these devices.

As we can see, one hour of the M608 printer being in Ready state uses the same amount of energy as being sleep mode for 56 hours. This shows that shaving off a few seconds between in idle times per job would lead to huge energy savings.

6.1.1 How the simulation work.

The simulation works by calculating how much energy is used between each print jobs. Using the symbols set out in section 3.2.5 it operates using the following logic.

$$Energy\ used\ between\ job = \begin{cases} S * B, & If\ S < timeout \\ (B * timeout)\ + E + ((S - timeout) * C1), & x \geq 0 \end{cases}$$

So for example, we know the time between two jobs is 500 seconds and the time out is 300 seconds. As time between jobs is greater than the timeout, this means the device goes to sleep so we would calculate the amount of energy used in idle mode plus the amount of energy used in the remaining sleep mode and the energy needed to wake up the device. Say this job is done on the on the M608 printer. We know being in idle for one second costs 0.0125 watts, a second in sleep mode is 0.0002 watts and wake up is 6 watts.

$$(300 * 0.0123) +\ 6 + \big((500 - 300) * 0.0002\big) =\ 9.73 \text{ watts used between jobs.}$$

Each print job is updated with the energy used between it and the next job and it notes if the device stayed in idle mode or if the device went to sleep. It is also noted if the device did not go to sleep for that long. I have defined this as an inefficient sleep as waiting in idle mode would of used less energy (time in sleep mode < threshold time).

6.2 Current policies

I decided that I would first test some of the current polices that I have come across during my research to see what a good starting point for policy modification and creation would be.

6.2.1 Oracle

Theoretically, the best possible result should be retrieved from an energy policy would be the oracle policy. Instead of having idle periods where energy is wasted happens, the device would turn its self-off as it knows the cost of switching from sleep to active is cheaper than staying in idle. Oracle requires the knowledge of the time between the next jobs in order to work, which be difficult to know in reality. However, it does gives us a goal that we can work towards.

Oracle				
Printer	Total energy used Watts	Stayed in idle	Went to sleep	Inefficient sleep
95	113662.31	5096	4983	1
96	204377.97	67158	9227	29
123	28433.93	3447	2491	0
133	22312.02	2398	1817	0
Total	**368786.2**	**78099**	**18521**	**30**

6.2.2 Energy Star Timeout

The energy star policy is the standard for many devices. The static timeout for these devices are typically around the 10 minutes mark so I have decided that I am going to set the results of this policy to be what I will comparing my all-future policies against.

Energy Star (10 minutes)				
Printer	Total energy used Watts	Stayed in idle	Went to Sleep	Inefficient sleep
95	155331.64	6034	3362	684
96	262127.82	68982	3964	3465
123	45959.34	3596	1923	419
133	35001.07	2522	1430	263
Total	**498419.9**	**81134**	**10679**	**4831**

From what we can see, out of the 96647 print jobs 84% of the time the printer was in idle mode when the next print job occurred and 16% of the time the device was in a sleep mode. Form setting the time to be 10 minutes there was not many inefficient sleeps.

From this point onwards all polices will be compared to the results gotten from 10 minutes to see how many times that policy it was more efficient or less efficient in terms of energy used.

6.2.3 Threshold

Setting the timeout to be equal to the threshold time produced the following.

Policy	TEU Watts	Idle	Sleep	Inefficient sleep	Jobs saved more	Jobs saved same	Jobs saved less
Threshold	475032.1	78098	11954	6595	15513	78098	3036

Compared to the ten-minute energy policy this used less energy but went to sleep more often leading to more cases where the device went to sleep just before it was needed. On 15513 occasions, the printer used less energy than the standard 10-minute timeout. The answer for this was simply going to sleep earlier when the time between both jobs is rather large. Even inefficient sleep could save move energy; it is just in this case the printer sleep faster than Timeout Inefficient sleeps, saving more energy.

6.2.4 Exhaustive search

Exhaustive search was run with times between 0 to 30 minutes to find a standard optimal time.

The following results is the combined amount of energy used by each printer.

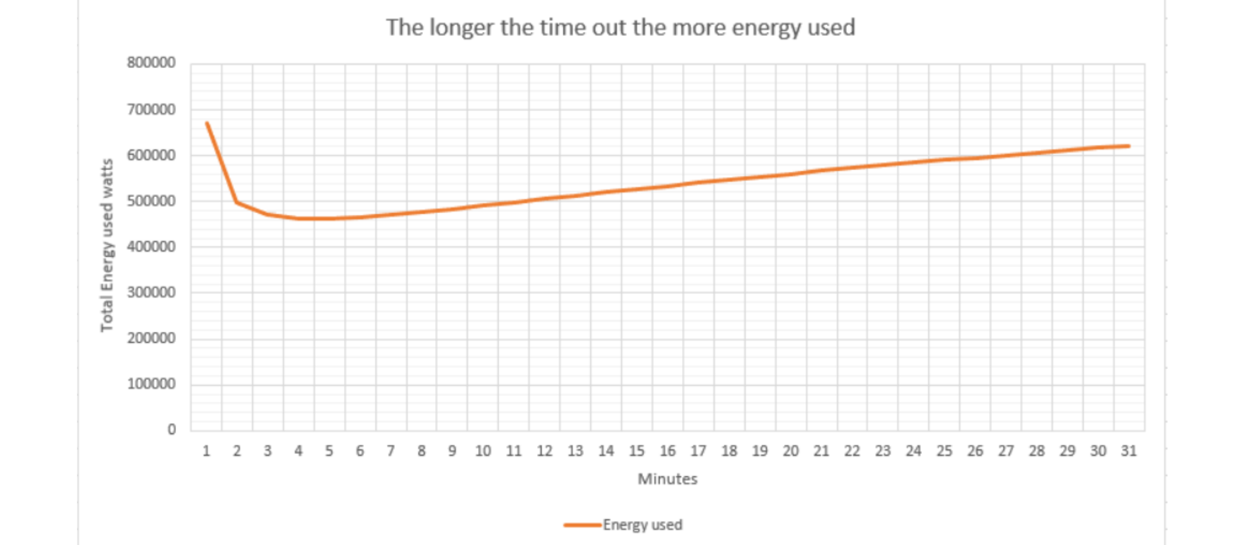


Figure 16: Exhaustive search results by total energy used

From this graph, we see that the amount of energy used increases the longer timeout is. Suggesting that a short timeout is best. However, the graph also shows that we must be careful, too short of timeout could cause the device to waste more energy than is needed. Like Oracle, we can only run this policy due to the fact we have a dataset of already done print jobs.

The following is the results for the time that used the less energy (4 minutes).

Policy	TEU Watts	Stayed in idle	went to sleep	Inefficient sleep	Jobs saved more	Jobs saved same	Jobs saved Less
Timeout 4 minutes	4628878.6	68995	14182	13470	15652	768995	12000

What we can see from this result that it saved a lot of energy however caused a lot of inefficient sleep.

6.2.5 L shape

Policy	TEU Watts	Stayed in idle	Went to sleep	Inefficient sleep	Jobs saved more	Jobs saved same	Jobs saved Less
L 0.1w	461305	68266	14010	14371	16897	68266	11484
L 0.5 w	462145.2	69101	13831	13715	16767	69101	10779
L 0.9 w	465955.2	70241	12845	13561	16597	70241	9809

The results retrieved from L shape showed a lot of energy saved but increased the number of inefficient sleep by over 3 times compared to the 10 minute timeout. When looking at the more efficient jobs it seemed too made up be jobs where the time between jobs was greater than 10 minutes or just over the threshold times and the L shape had set the time out time to be zero. This meant that the device just went to sleep immediately meaning no energy was spent in idle mode. On the other side, the less efficient job where caused when by the policy decided that the timeout should be zero when the time between jobs was less than threshold.

6.3 The problem discovered from running the current policies

If we set the printer to go to sleep too fast, then we are likely to face an issue where the printer goes to sleep just before a print job leading to energy wastage. In this case remaining in idle would have been more efficient.

If the set timeout is large but the time between jobs is larger then energy is being wasted staying in idle mode and then going to sleep to then wake up later. In this case, going to sleep from the get-go would have been more efficient.

If the set timeout is larger than the breakeven, but we go to sleep just before the sleep period then we have just spent energy equal to two wake ups as opposed to one wake up and (sleep mode energy * time difference).

6.4 Polices based off analysis

In this section I will be going over some of the polices that I have created that use information discovered in my analysis. All results will be compared to the standard ten minutes time out policy to see how effective they are.

6.4.1 Time

First, I decided to investigate if just using different timeouts based on typical timeouts form the previous years would produce better results. I decided to test setting the timeout to be equal to the median time between jobs and testing setting the timeout to be equal to the upper quartile time between jobs.

6.4.1.1 Corresponding CSV Read in policy

These polices work by getting information from the current print job and then using said information to find its corresponding recommend timeout value in a CSV file. For example if are using the policy based on hour and the user prints something off at the 08:39:56 we would get the hour of this job (8 o'clock) and then search the appropriate table that is made up of recommended timeouts for each hour. We would get the value that corresponds to the 8 o'clock and use that as the timeout.

Each policy has its own set of values it can use. Hour has 24 options, weekday has 168 option (24 * 7) and so on.

Corresponding Hour

$$Time\ out = CorrespondingTimeOut(Job\ Hour)$$

Policy	TEU Watts	Stayed in idle	Went to sleep	Inefficient sleep	Jobs saved more	Jobs saved same	Jobs saved Less
Hourly median	499329.2	45149	17004	34494	18567	45149	32931
Hourly UQ	642246.6	68706	11095	16846	8338	65766	22543

Corresponding Weekday and Hour

$$Time\ out = CorrespondingTimeOut(Job\ Hour\ \&\ Job\ Weekday)$$

Policy	TEU Watts	Stayed in idle	Went to sleep	Inefficient sleep	Jobs saved more	Jobs saved same	Jobs saved Less
Weekday Hour Med	499869.3	45218	16993	34436	18471	45218	32958
Weekday Hour UQ	485778.2	67826	13825	14996	14518	67455	14674

Corresponding Month and Weekday and Hour

$$Time\ out = CorrespondingTimeOut(Job\ Hour\ \&\ Job\ Weekday\&\ Job\ Month)$$

Policy	TEU Watts	Stayed in idle	Went to sleep	Inefficient sleep	Jobs saved more	Jobs saved same	Jobs saved Less
Month Med	518772.5	45359	16640	34648	17957	45280	33410
Month UQ	544338.3	67828	13005	15814	12785	66933	16929

6.4.1.2 Read in time discussion

These results show that the more in-depth I made my policy the worse energy saving became. Investigation discovered that

> Using the median value seemed to have an issue that it was just sort of the actual timeout in most cases. I expected this to happen, as the median is the middle point and tells us that it only going to benefit half of the jobs. However, many of the times where the median did better is when the difference between jobs is so large that it's greater than energy star 10 minutes and the low timeout meant the device went to sleep faster.
>
> Using the upper quartile, the opposite happened. Longer timeouts outs meant that the device stayed in idle mode for longer so caught more user prints. However, there was a lot more occasion where the device kept itself in idle mode for a longer time than the standard energy star so in some cases more energy was wasted.
>
> The month policy used a lot more energy. This is primarily due to the changes in the semester and that a flat month policy to cover those events would inefficient.

This suggest that I need to build a controller to help limit faults. We know that setting the timeout to be equal to the breakeven is somewhat more efficient than the ten minutes energy star policy. The only time in which it fails to be better is when the time between jobs is less than the 10 minutes but more than breakeven.

6.4.1.3 Controller corresponding policy

This is a somewhat modified read in corresponding policy that detects if the predicted given time is less than the printer's threshold time then set the timeout to be equal to threshold instead. If the time is truly less than the threshold then this is not going to influence the energy saving at all because most jobs will already be happen in less time, this would catch a few outliers. Another shortcoming of the previous read in corresponding policy is when the predicted timeout is long but the actual time between jobs is a lot longer meaning a lot of energy is wasted in idle. To combat this the following policy has been created.

$$Timeout = \begin{cases} Threshold\ time, & If\ CorrespondiUQ < Threshold \\ CorrespondingMedianTime(X), & Otherwise \end{cases}$$

If the upper quartile is greater than the threshold it suggest we may want to go to sleep sooner so the median value will be used instead.

After running test using the new policy I got the following results.

Policy	TEU Watts	Stayed in idle	Went to sleep	Inefficient Sleeps	Jobs saved more	Jobs saved same	Jobs saved Less
CcHourly	470659.6	77387	12190	7070	15587	77387	3673
CcWeekday	470120.7	77310	12188	7149	15545	77310	3792
CcMonth	475814	76317	12546	7784	15312	76238	5097

This new policy has saved more energy than the standard and threshold time as timeout. However, it does increase the inefficient sleep by roughly 3000.

To do further improvement I wanted to see where these policies failed to be more efficient than the energy star algorithms. These failures were primarily during the busy hours where the device shut down just it was needed again.

When investigating the hour version of this policy it seemed 883 of the inefficient sleeps were done in slow hours and 6187 were done during busy hours.

6.4.1.4 Controlled corresponding modified.

As we saw in the previous policy, most of the inefficient sleeps where caused by the device going to sleep too early during busy hours. I modified the last policy to double timeout if the read value is believed to be one of a busy period.

$$Timeout = \begin{cases} Threshold\ time * 2, & If\ CorrespondiUQ < Threshold \\ CorrespondingMedianTime(X), & Otherwise \end{cases}$$

Policy	TEU Watts	Stayed in idle	Went to sleep	Inefficient sleep	Jobs saved more	Jobs saved same	Jobs saved Less
Ccv2Hourly	518391	84277	2910	9460	5009	80978	10660
Ccv2Weekday	516879	84244	2907	9496	5314	80943	10390
Ccv2Month	505646	83381	3304	9926	7611	80648	8388

The result gotten from this policy is worse than the standard in terms of energy used. What is interesting about these results is that when we look at the result gotten from the using Month it clearly shows that the more the device goes to sleep then the less total energy used.

6.4.2 By Devices

As previously established, each device has very different usages meaning what works on one device may not work on another. I am going to see if giving each printer its own set of values can increase the efficiency of the printers.

6.4.2.1 Corresponding policy

6.4.2.1.1 Hourly

$$Time\ out = CorrespondingTimeOut(Job\ Hour\ \&\ Device)$$

Policy	TEU Watts	Stayed in idle	Went to sleep	Inefficient sleep	Jobs saved more	Jobs saved same	Jobs saved Less
Hourly median	528299.6	46008	16137	34502	17622	45990	33035
Hourly UQ	642246.6	68706	11095	16846	8338	65766	22543

6.4.2.1.2 Weekly

$$Time\ out = CorrespondingTimeOut(Job\ Hour\ \&\ Job\ Weekday\ \&\ \ Device)$$

Policy	TEU Watts	Stayed in idle	Went to sleep	Inefficient sleep	Jobs saved more	Jobs saved same	Jobs saved Less
Week med	540986.5	46096	16024	34527	17068	46016	33563
Week UQ	664481.1	68724	11038	16885	8449	65767	22431

6.4.2.1.3 Monthly

$$Time\ out = CorrespondingTimeOut(Job\ Monthly\ \&\ Job\ Hour\ \&\ Job\ Weekday\ \&\ \ Device)$$

Policy	TEU Watts	Stayed in idle	Went to sleep	Inefficient sleep	Jobs saved more	Jobs saved same	Jobs saved Less
Month Med	717240	46570	15005	35022	15072	45894	35681
Month UQ	911406.2	68619	9975	18053	7951	65196	7951

As we can see this caused some poor results, which was surprising as we saw in chapter 5.3 that the different printers had different usage. All results used more energy and cause more inefficient sleep.

6.4.2.2 Controlled Corresponding printer

When applying the policy created in the previous section to work with printers, we retrieve the following.

$$Timeout = \begin{cases} Threshold\ time * 2, & If\ CorrespondiUQ(X + Printer) < Threshold \\ CorrespondingMedianTime(X), & Otherwise \end{cases}$$

Policy	TEU Watts	Stayed in idle	Went to sleep	Inefficient sleep	Jobs saved more	Jobs saved same	Jobs saved Less
ccPrinter Hour	465442.3	76756	12460	7431	15417	76738	4492
ccPrinter Weekday	470120.7	76692	12382	7573	14883	76612	5142
ccPrinter Month	646111.5	76845	11840	7962	13076	76169	7402

The policy that just used hours produced the best result in terms of energy used and inefficient sleeps (though it's more than the standard).

6.4.3 By Term time

We discovered in chapter 5.3 that the term time can impact the device usage, as we have seen the last few polices that straight up use month also produce a poor result it suggest that a policy around term time would be beneficial. Previous polices have clearly shown that the polices that read in values produce bad results. In this section I have went straight into using the controller policy.

6.4.3.1 Controlled Corresponding printer

Breakeven - Check Hourly							
Policy	TEU Watts	Stayed in idle	Went to sleep	Inefficient sleep	Jobs saved more	Jobs saved same	Jobs saved Less
ccHour	464764.7	76163	12672	7812	15659	76160	4828
ccWeekday	468103.9	76150	12682	7815	15556	76107	4984
ccMonth	470103.9	76070	12689	7888	15533	76045	5069

Like seen in previous results, total energy used better than standard policy to produce sleeps that are more inefficient.

6.4.4 By Users

Previous policies have focused on the device and not the people using it. This following section focus on using the number of print jobs a user has committed in their current session to set an appropriate timeout. Total prints will represent the number of print job a user has done in the past hour.

6.4.4.1 Policy User history

$$CTOUQ = CorrespondingTimeOutUQ$$

$$TimeOut \,\&\, LastTimeout = \begin{cases} CTOUQ(Total\ Prints), & If\ CTOUQ(Total\ Prints) > (LastTimeOut - Time\ between\ last\ job) \\ (LastTimeOut - Time\ between\ last\ job, & Otherwise \end{cases}$$

This ensure that the user with the longest time between takes priority.

Breakeven – Device							
Policy	TEU Watts	Stayed in idle	Went to sleep	Inefficient sleep	Jobs saved more	Jobs saved same	Jobs saved Less
User	519110.8	80274	10329	6044	9642	77779	9226

The results retrieved from this policy is worse than the standard.

6.4.4.2 Policy User history version 2

From the poor results retrieved from the previous policy, I decide to make a similar policy based of my controlled corresponding policy

$$Timeout = \begin{cases} Threshold, & If(CTOUP(Total\,prints) < Threshold \\ CTOmedian(Total\,prints), & Otherwise \end{cases}$$

Policy	TEU Watts	Stayed in idle	Went to sleep	Inefficient sleep	Jobs saved more	Jobs saved same	Jobs saved Less
Breakeven User	519464.1	82624	9868	4155	9569	80129	6949

6.4.4.3 Policy User history version 3

A modified version of the first user policy. We do not know is a user has printed off one document is going to print another so it may be safer to set the median

$$TimeOut = \begin{cases} CTOM(Hour), & If\ Total\,prints = 1 \\ Threshold, & If(CTOUQ(Total\,Print) < Threshold \\ CTOMedain(total\,print), & OtherWise \end{cases}$$

More in depth user policy

Breakeven – Device							
Policy	TEU Watts	Stayed in idle	Went to sleep	Inefficient sleep	Jobs saved more	Jobs saved same	Jobs saved Less
User	473826.9	76957	12136	7554	15598	76957	4092

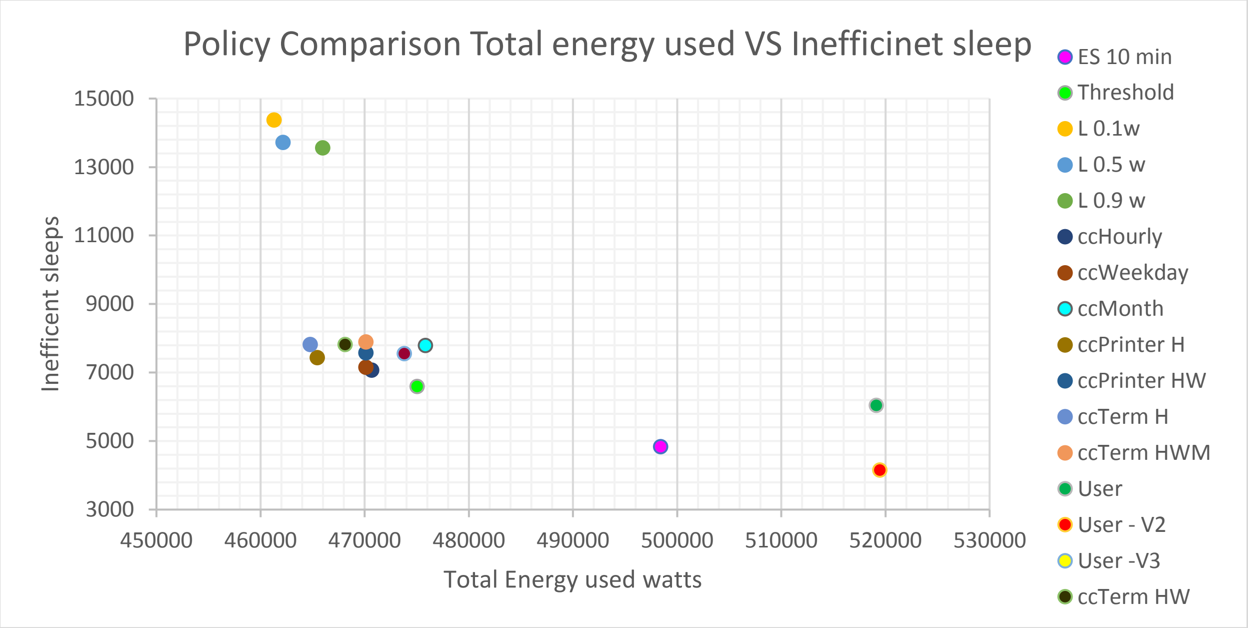
6.5 Summary

Throughout this section I've developed many different polices based off analysis that had been done in Chapter 5. The aim of this dissertation was to create a policy using data analysis that could have produced a result that optimized a printer in term of energy used and un-optimized shut off.

The results include in this table are from policies who's jobs tended to the same or saved more than the ES 10 minute timeout.

Policy	TEU Watts	Stayed in idle	Went to sleep	Inefficient sleep	Jobs saved more	Jobs saved same	Jobs saved Less
ES 10 min	**498419.9**	**81134**	**10679**	**4831**			
Oracle	**368786.2**	**78099**	**18551**	**0**			
Threshold	475032.1	78098	11954	6595	15513	78098	3036
L 0.1w	461305	68266	14010	14371	16897	68266	11484
L 0.5 w	462145.2	69101	13831	13715	16767	69101	10779
L 0.9 w	465955.2	70241	12845	13561	16597	70241	9809
Created Policy							
ccHourly	470659.6	77387	12190	7070	15587	77387	3673
ccWeekday	470120.7	77310	12188	7149	15545	77310	3792
ccMonth	475814	76317	12546	7784	15312	76238	5097
ccPrinter H	465442.3	76756	12460	7431	15417	76738	4492
ccPrinter HW	470120.7	76692	12382	7573	14883	76612	5142
ccPrinter HWM	646111.5	76845	11840	7962	13076	76169	7402
ccTerm H	464764.7	76163	12672	7812	15659	76160	4828
ccTerm HW	468103.9	76150	12682	7815	15556	76107	4984
ccTerm HWM	470103.9	76070	12689	7888	15533	76045	5069
User – V1	519110.8	80274	10329	6044	9642	77779	9226
User - V2	519464.1 ,	82624	9868	4155	9569	80129	6949
User -V3	473826.9	76957	12136	7554	15598	76957	4092

This the resulting scatter graph based off the result retrieved from each policy (oracle excluded to it not really being feasible).



The point produced by energy star is in bright pink. As we can see a majority of the policy tested and created in this project easily reduced the amount of energy used a majority saved at least over 30000 watts. Orcale showed the max that we could save is 120000 watts meaning most policy where able to at least get ¼ of the maximum possible energy saved. However, what we can clearly see that the policies that saved energy tended to increase the inefficient sleeps. Most polices increased inefficiencies by 3000 (a 60% increase. The L shape algorithms tended to save the most energy but increase inefficiency by 250%. The one policy (User – v2) that got a worse energy saving than ES 10 minutes was the only policy to get less inefficient sleeps than ES.

These findings are tell me that that if you want to optimize a policy then unfortunately you only have two options

Optimization in terms of fewer delays to a user

Optimization in terms of less energy used.

Though my policies may not produce an optimal result, I believe that I have found a better policy that find than the current standard. My Controlled Corresponding Printer hour policy and My Controlled Corresponding Term Time hour policy both produced a an energy saving that produced a that only 1% more than the energy used by L shape 0.1 and cause half of the inefficient energy usage. This improvement over L shape suggest that data analysis has been useful in creating an energy policy. The best results and polices seemed to be those that require less parameters to find the corresponding value. It also seemed that polices that revolve around the device produced better results than those base on the user.

6.5.1 What’s holding savings back

A big problem is that it is difficult for the device to truly know when it is next going to be used. As if the user has done 2+ jobs then you could roughly predict when the next print job is going to be. The real problem is determining if a user who has printed off one document is going to print off another.

Just using the printer information, when we detect that a new user has printed something and there has not been a print job in a while, we can respond in two ways.

Do we keep the device in idle mode?

> If the user only prints off one document, then any amount of time spent in idle mode is wasted if the printer is not going to be used again.

Do we put the device into sleep mode?

> If a user is going to print off another document but the device is asleep then a delay will be caused as the device wakes up, not to mention that energy will be wasted.

To review, in all scenarios energy is going to will wasted at some point.

I did some analysis to see if there were times where one scenario that was more common than the other was. I found that in all situation both where too evenly matched (near 50\50) to confidently make a real decision.

6.5.2 Solution

The original simulation would have allowed me to run user logon data alongside the printers. This meant I would have been able to track when a user logons and prints. I believe that this would have been the key to solving this problem.

When looking at users log on and prints it was discovered that a large majority of all print jobs occur within 15 minutes of a user logging on. This gave me an idea of a policy that would listen to when a user had logged on, how many people where currently logged on and who had recently logged of. If we know that a user has just logged off, then there is no point staying in idle for them. If a person has just logged on, then it may be worth increasing the current idle time by a bit as another job could be coming soon. Appendix 2 goes into my analysis of this more.

7 Conclusion

7.1 Satisfaction with Aims and Objectives

One: Research and summarize energy policies

This objective was fulfilled through the research in section 3.2 where I discuss existing policies and how they worked.

Two: Perform data analyses to find trends in printer usage

This was fulfilled through Chapter 4 where I performed an in-depth data cleaning to get the data prepared and Chapter 5 where I analyzed several aspects of the data.

Three: Implement and simulate different policies

Though there was a bit of a hic-up, a simulator was eventually created so that polices could be implemented then simulated as seen in Chapter 6.1.

Four: Modify policies to discover potential avenues for optimization

Whilst not I was not able to create as many policies as I would have liked, some polices have been created and modified using information gained off my analysis, as seen Chapter 6.

Five: Analyze and compare different policies to find the optimal energy-performance trade-off

In Chapter 6.5 a range of police where compared to one another to see which was the most efficient and possible fixes that could be implement to find a better solution.

7.2 Effectiveness of CRISP-DM

As stated in Chapter 2, I used the CRISP-DM model. I believe following this model greatly helped my dissertation as it gave me a guide as to what I could next meaning that I was always moving forward. It also ensured that I was constantly going back to my aims and objectives to review what I had recently done. This constant self-reviewing meant I had many opportunities to step back and see what went well and what I could do to improve.

7.3 Personal development

I felt that I have developed and a lot over this project.

- I've learned how to efficiently research into a topic as seen in chapter 4
- I've gained an understanding into the world of data analysis and how to do it properly
- I feel like I'm better at problem solving by solving issues that have occurred in my dissertation.
- I've developed skills in R, R-studio, ggplot2 and tidyverse

7.4 What could have gone better?

In the end, I had to create my own simulation in order to get results. This was a sudden shift in the dissertation that I had to deal with and it did cause some issues, mainly time. I was able to get something up, running but it was rushed, and I will admit that it’s not the most amazing bit code in world. If I had more time to work on the simulation, then I could have created a more complex simulation that would could allowed me to develop more advanced policies.

7.5 Future work

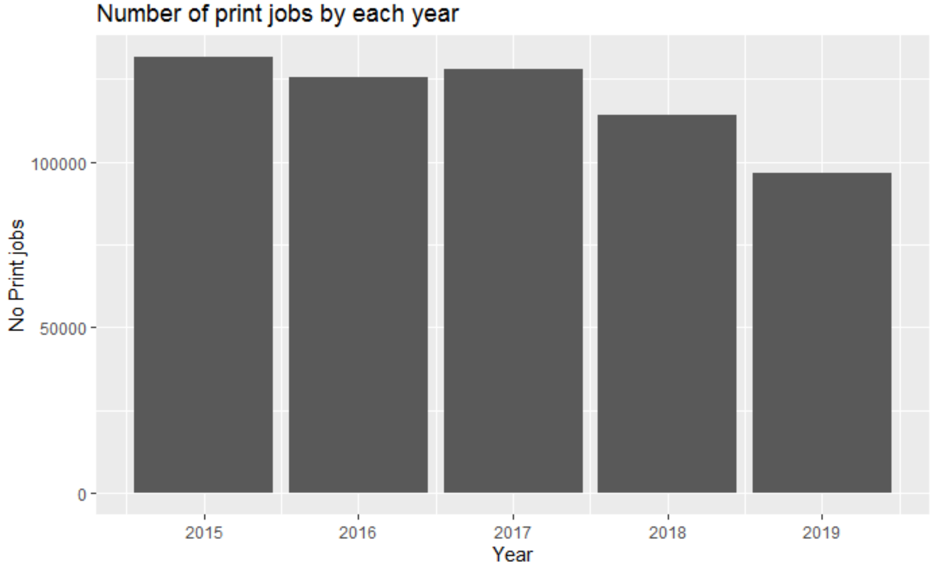
There were many areas that I did analysis in but did not get a chance to convert to a policy due to different reasons. Like stated in chapter I believe that 6.5.2 , believe user login information could be useful in generating a better solution.

Appendix 3 included some work that was started looking into doing some basic pattern analysis. It could be used to help resolve the issue I ran into with predicting the next users' jobs.

It would also be interesting to see if findings and policies used in this project could be directly applied to another device with different usage rates to discover how effective they would be. This work could be further developed to a new policy that could be universally applied to other devices that could optimize their energy usage.

Appendix 1

Investigation of year



Bar chart of number of prints per year

- Firstly, analysis of the number of print jobs in each given year was done.
- I was expecting to see a decline in print jobs with each year and the resulting graph meets my expectations.
- As the years go one, the number of print jobs start to decline. The increasing number of subjects adopting digital coursework's submissions may cause a possible explanation for this. Another explanation is this could be due a social push to combat paper wastage, a push that the university has started to follow by. As people are becoming more conscious to what they are printing out then I will predict that the total for 2020 would also follow this pattern of decrease.
- As the printer policy will be applied to future years then it's likely that that whatever is made will have to deal with less print and therefore a higher average time between print jobs. This suggest that it's now more important to work out the ideal idle period in order to not waste energy.
- Just looking at the year will provide us with nothing to concrete, there are many other factors that could come into play, such as months, days and the hours of the day

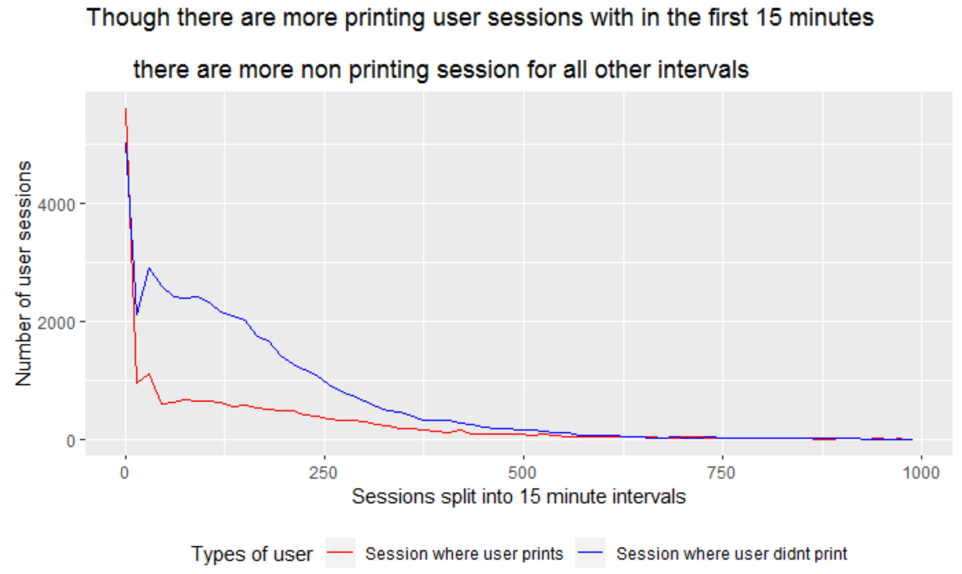
Appendix 2

A user's logins and prints

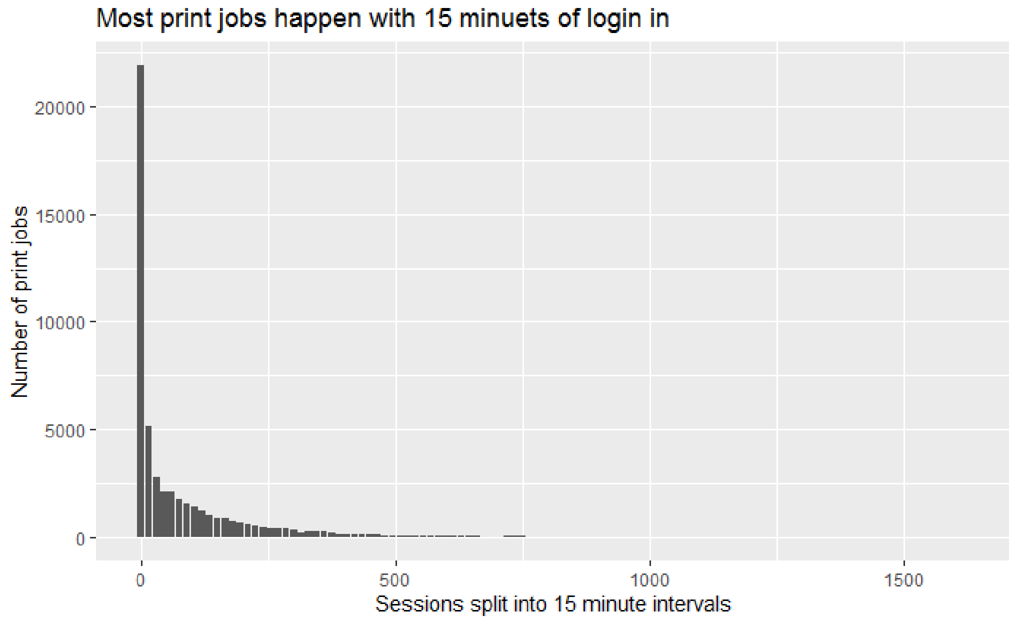
The last few chapters have had me investigating the usage of a device to try and create a policy that would optimism energy efficiency, in this chapter I wanted to look at the users as an individual and how they tend to use the device.

First, I wanted to look at the differences between those who printed off a document and those who didn’t. Given the login data for 2019 I matched session to print jobs and retrieved the following.

The number of user sessions with no print jobs is 45414 and the number of sessions where the user printed something out was 19493. If we round both these number to the nearest 500, we get a 7:3 ratio. Meaning that for every person that logs on then there is a 30% chance the user will print. Though the ratio is waited towards those who don’t print I decide to see if there were as any patterns that could help me create a policy.

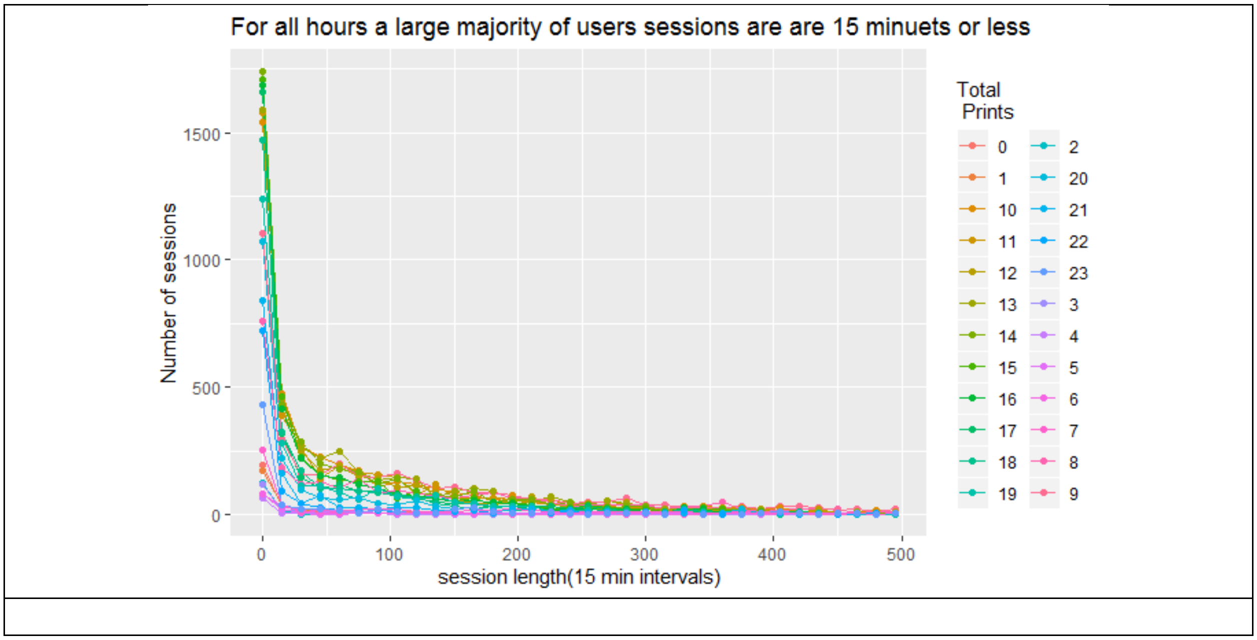


As there was a lot of user session of various lengths, I decided to a group the user together by session length of 15 minutes. What this graph tells us is that in cases if you were to pick a user based on session length then its most likely they will be a non-printer. However, as you can see that when it comes to user session that are between 0- and 15-minutes user session who print overtake the sessions where a user didn’t. The percentages of this are 47% non-printer to 53% printer users. The graph also shows that most users are logging in printing something out then going as most print users are between 0 – 15 minutes.



When we count how many jobs where done in each interval, we discover that 43% of jobs where done within 15 minutes of a user logging in, again suggesting that people are logging in and then printing something out almost immediately. When calculated the median it was discovered that the midpoint of the user session length is falls at around 90 minutes.

As we now know a user who wants to print is likely to do it within minutes of logging then if we can get an idea of when people are logging in, we could use that information to help make predictions. Upon looking at the most common user session by hour its clear to see that 15 minutes are the most common throughout all hours.



Originally, I had an idea to create a policy that would detect when someone has logged so it would turn its self on however if we account for how many people don't print then we would have a policy that would be only 12% effective. Meaning that it may be more harmful than helpful if there is just a single user logged on as more often than that it would mean more pointless starts and shutdowns which would waste energy and would wear down a device faster.

But there might be something to this. We know that if a person logs on then and they are wanting to print something off then they are more likely to print something off sooner than later, so if a printer is about to exceed its time out time but someone logs in then it may be beneficial to just increase the time out by a few minutes just in case the user is going to print something out. Though this again would only be 12% effective it may be beneficial as the energy used for a few more minutes in idle mode would still be less energy used by a wake up.

This finding then gave me the following ideas.

If a user is the only person in a room printing something off, then if we could work out how likely that person is going to print off again. If that could be figured out, then in may be interesting to see if there is any difference in how long a user wait till they print again.

As we have all the logins, we it may be worth seeing if the number of people in a cluster room influences the number of print jobs.

Appendix 3

This is a section on pattern recognition that I was working on in my last cycle. I decided that I didn't have enough time to finish this section which is a same as I feel like there is a lot of useful information that could be gained. It certainly would help solve my problem with working out if a user just going to print off one job.

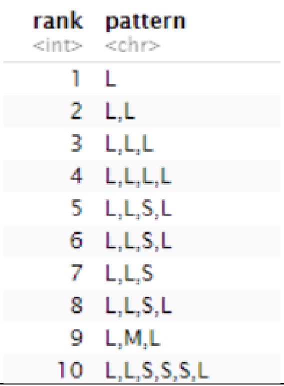
If we can work patterns common in printing, then we could use that to create a policy that tries to predicate what the next print job could be given the history. To do this I broke the times between jobs for the years 2014 - 2018 to be

S = time difference less than 5 minuets

M = Between 5 and 10 minuets

L = Greater than 10 minuets.

After ranking how often each print pattern with each hour I then compared all the hours together to create this ranking list.



rank <int>	pattern <chr>
1	L
2	L,L
3	L,L,L
4	L,L,L,L
5	L,L,S,L
6	L,L,S,L
7	L,L,S
8	L,L,S,L
9	L,M,L
10	L,L,S,S,S,L

Top 10 Hourly print patterns

As we see in the top 10, most popular print pattern is just one print job.

References

Statista. 2020. *Iot: Number Of Connected Devices Worldwide 2012-2025 | Statista*. [online] Available at: <https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/> [Accessed 10 April 2020].

Energysavingsecrets.co.uk. 2020. *Does Having Appliances On Standby Use Power?*. [online] Available at: <http://www.energysavingsecrets.co.uk/does-appliances-standby-use-power.html> [Accessed 10 April 2020].

Francis, J., 2014. *Understanding And Performing Standby Power Measurements*. [online] Power Electronics. Available at: <https://www.powerelectronics.com/technologies/power-electronics-systems/article/21861639/understanding-and-performing-standby-power-measurements>.

Commission Regulation (EC) No 1275/2008. 2008 implementing Directive 2005/32/EC of the European Parliament and of the Council with regard to ecodesign requirements for standby and off mode electric power consumption of electrical and electronic household and office equipment (Text with EEA relevance) *OJ L 339, 18.12.2008, p. 45–52 (BG, ES, CS, DA, DE, ET, EL, EN, FR, IT, LV, LT, HU, MT, NL, PL, PT, RO, SK, SL, FI, SV)*
Special edition in Croatian: Chapter 13 Volume 054 P. 144 – 151

Shrestha, P., 2020. *Wasted Energy 'Costing Businesses £60M A Year In Unnecessary Bills' - Energy Live News*. [online] Energy Live News. Available at: <https://www.energylivenews.com/2020/01/22/wasted-energy-costing-businesses-60m-a-year-in-unnecessary-bills/>.

James, P. and Hopkinson, L. (2009). Sustainable ICT in Further and Higher Education. [online] Ictliteracy.info. Available at: http://www.ictliteracy.info/rf.pdf/rptgreenictv1.pdf [

Microsoft Corporation, 2020. *Microsoft Excel*, [online] Available at: https://office.microsoft.com/excel.

Microsoft Corporation, 2020. *Microsoft PowerBi*, [online] Available at: <https://office.microsoft.com/powerBi>

R studio 2020. [online] R studio Available at:<https://rstudio.com/>

TidyVerse 2020. [online] Available at:<https://www.tidyverse.org/>

Ggplot 2020. [online] Available at: <https://ggplot2.tidyverse.org/>

Forshaw, M., McGough, A. and Thomas, N. (2015). HTC-Sim: a trace-driven simulation framework for energy consumption in high-throughput computing systems. CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE, pp 1-32

Jensen, K., 2012. A Diagram Showing The Relationship Between The Different Phases Of CRISP-DM And Illustrates The Recursive Nature Of A Data Mining Project.. [image] Available at: <https://commons.wikimedia.org/wiki/File:CRISP-DM_Process_Diagram.png> [Accessed 10 April 2020].

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Shearer, C. and Wirth, R. (2019). CRISP-DM 1.0. pp 13-15

Resnik, D. 2000. Statistics, Ethics, and Research: An Agenda for Education and Reform. Accountability in Research 8: 163–188.

Kassner, M., 2017. 5 Ethics Principles Big Data Analysts Must Follow. [online] TechRepublic. Available at: <https://www.techrepublic.com/article/5-ethics-principles-big-data-analysts-must-follow/>

Shamoo, A. and Resnik, D., 2014. Responsible Conduct Of Research. Oxford: Oxford University Press.

Orbit. 2020. *AREA 4P Framework | Orbit RRI*. [online] Available at: <https://www.orbit-rri.org/about/area-4p-framework/#1467057166816-da75e065-e481> [Accessed 10 April 2020].

Shamoo, A. E., and Resnik, D. B. 2006a. Ethical Issues for Clinical Research Managers. Drug Information Journal 40: 371–383.

RCR. 2020. Data Analysis. [online] Available at: <https://ori.hhs.gov/education/products/n_illinois_u/datamanagement/datopic.html>

Smeeton, N., Goda, D. 2003. Conducting and presenting social work research: some basic statistical considerations. Br J Soc Work, 33: 567-573.

D. J. Harriss, A. Macsween, G. Atkinson (2018). Standards for Ethics in Sport and Exercise Science Research: 2018 Update

Fernandez, S., 2019. Data Cleansing Combating The Cost Of Bad Data. [online] Habiledata.com. Available at: <https://www.habiledata.com/blog/data-cleansing-combating-the-cost-of-bad-data/> [Accessed 7 May 2020].

Etlinger, Susan. 2014. What do we do with all this big data [Video file]. Retrieved from https://www.ted.com/talks/susan_etlinger_what_do_we_do_with_all_this_big_data

D. J. Harriss, A. Macsween, G. Atkinson .2018. Standards for Ethics in Sport and Exercise Science Research: 2018 Update

Gottschalk, L. A. 1995. Content analysis of verbal behavior: New findings and clinical applications. Hillside, NJ: Lawrence Erlbaum Associates, Inc

Edge Analytics. 2020. *Edge Analyticsdata Analysis & Big Data Case Studies | Edge Analytics*. [online] Available at: <https://edgeanalytics.co.uk/case-studies.php/>

venkatesh ,2013, India Study Channel.. *Energy Wastage And Conservation In Daily Life Important Days*. [online] Available at: <https://www.indiastudychannel.com/resources/160848-Energy-wastage-conservation-daily-life.aspx>

Energy.gov. 2020. Energy Efficiency Policies And Programs. [online] Available at: <https://www.energy.gov/eere/slsc/energy-efficiency-policies-and-programs>

Energystar.gov. 1992. ENERGY STAR | The Simple Choice For Energy Efficiency. [online] Available at: <http://www.energystar.gov>.

PETERSON, J. L. AND SILBERSCHATZ, A. 1985. Operating System Concepts. 2nd ed. Addison-Wesley Series in Computer Science. Addison-Wesley Longman Publ. Co., Inc., Reading, MA.

C.-H. Hwang and A. C. Wu. . 1997. A Predictive System Shutdown Method for Energy Saving of Event-Driven Computation. In International Conference on Computer-Aided Design, pages 28– 32, Srivastava, M. B., Chandrakasan, A. P. and Brodersen, R. W. (1996) Predictive system shutdown and other archi-tectural techniques for energy efficient programmable computation.IEEE Trans. Very Large Scale Integr. Syst.,4, 42–55.

A. Karlin, M. Manasse, L. McGeoch, and S. Owicki. 1994. Competitive Randomized Algorithms for Nonuniform Problems. Algorithmica, 11(6):542–571.

Weisstein, Eric W , 2020. "Exhaustive Search.", [online] Available at: <https://mathworld.wolfram.com/ExhaustiveSearch.html>

E.-Y. Chung, L. Benini, and G. D. Micheli. 1999. Dynamic power management using adaptive learning tree. In International Conference on Computer-Aided Design, pages 274–279, 1999.

T. Simuni ˇ c,´ L. Benini, and G. D. Micheli. 2000 . Dynamic Power Management of Laptop Hard Disk. In Design Automation and Test in Europe.

Yung-Hsiang Lu, Eui-Young Chung, Tajana,, Simuni ˇ c,´ Luca Benini, 2008, Quantitative Comparison of Power Management Algorithms

Wikipedia D ,2020, Data cleansing [online] Available at: https://en.wikipedia.org/wiki/Data_cleansing#Error_event_schemal

Simran Kaur Arora , 2020, What is Data Analysis? Methods, Techniques & Tools [online] Available at: https://hackr.io/blog/what-is-data-analysis-methods-techniques-tools

Laser jet M608 , 2020 , [online] Available at:<http://www.hp.com/support/ljM608>

Laser jet M651 , 2020, [online] Available at:https://support.hp.com/gb-en/drivers/selfservice/hp-color-laserjet-enterprise-m651-series/5391209

DigitalVyda , 2018, 8 Ways to Clean Data Using Data Cleaning Techniques , [online] Available at:<https://www.digitalvidya.com/blog/data-cleaning-techniques/#>

John Dillard , 2020, The Data Analysis Process: 5 Steps To Better Decision Making , bigskyassociates , [online] Available at:https://www.bigskyassociates.com/blog/bid/372186/The-Data-Analysis-Process-5-Steps-To-Better-Decision-Making

Newcastle University logo, 2020 , [online] Available at: <https://blogs.ncl.ac.uk/alc/newcastle-university-logo/>