Towards student study space usage monitoring using Wi-Fi data

Dimitris Christodoulou

MInf Project (Part 1) Report
Master of Informatics
School of Informatics
University of Edinburgh

2021

Abstract

Finding an empty study space at the University of Edinbugrh's Main Library can be challenging at times, especially during busy periods such as exam seasons. Better information on the availability of study spaces can help students find a space more quickly and easily, and help the Library's management manage the available resources more effectively. When a device connects to a Wireless Access Point, an event is stored after personal information is removed. The data is simple - each event only includes its timestamp and the access point where it was recorded. Even though for this project data from the University of Edinburgh's Main Library is used, the outcome of this project can be adapted and used elsewhere in the University, in other Libraries or in other settings such as museums or conference centres.

The characteristics of the available data are explored and analysis is carried out to assess the possibility of the data revealing personal information, despite all personal identifiers and dates having been removed. A preprocessing pipeline is established, including noise removal, segmentation of the Wireless Access Point log into sessions and determining the primary device for each user.

A model is constructed that can predict whether a user is settled at some area of the building at any given time. The classification features are calculated within a window preceding the time for which the prediction is made. We propose a way to change the size of this window adaptively, depending on the volume of events. The model used is a Random Forest Classifier, selected after several candidate models were compared and their hyperparameters were tuned using cross-validation.

The results of the model are evaluated using a separate testing set in two ways. The first, is on a number of manually labelled examples. The second is by using the output of the model to build a classifier that can predict whether an access point is located at a study space or not, and then comparing against our knowledge of which access points are at study spaces.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Petros Papapanagiotou for his guidance and support throughout this project.

I would also like to thank my friends and family for supporting me during these challenging times.

Table of Contents

Chapter 1

Introduction

In the first chapter, we start by discussing the motivation behind this project and outlining its main goal (Section 1.1). Then, the general methodology we followed is described (Section 1.2) and the key contributions of the project are listed (Section 1.3).

1.1 Motivation and Goal

The Main Library is one of the most important buildings in the University of Edinburgh's George square campus, used by thousands of students and staff daily. It provides several facilities, including book collections, study spaces, workshops, printing equipment, meeting spaces and a cafe.

The University's student population has increased significantly in the last few years - increasing by approximately 25% in the last five years [12, 13]. This has resulted in increased demand for the University's facilities, including study spaces in the Main Library. Inability or long delays in finding an empty study space has been a common source of frustration among students [4].

The Library has already taken action to address these issues. In an attempt to alleviate the pressure on study spaces, they have increased the number of study spaces by 358 during the last four years [17]. Despite this, demand still outstrips supply. An effort has been made to improve the information that is provided to students on the availability of study spaces. For example, a simple traffic light system indicates how busy the library is overall at any given time. The work so far has been well received by students.

The overall goal of the project is to explore ways which existing data sources, most importantly, data from Wi-Fi Access Points can be used to provide insights regarding occupancy monitoring and the use of study spaces among other library facilities. The outcome of this project could be useful to both students who use the library and to the Library management. Better information can be provided to students about where exactly they can find an empty study space. Additionally, the Library Management will be able to gain more detailed insights about how many people are using study spaces, which study spaces and potentially monitor resource usage across the library.

More specifically, the key outcome in this report is the development of a model that can predict at any given time whether a user is settled at a particular area of the main library or not. This can be used to distinguish between students who are actively using a study space and those who could be moving, but are within the range of nearby access points. This could then be used to more accurately estimate the number of available study spaces in different areas of the main library.

As the data used is simple, the outcome of this project can also be adapted and used in a variety of other settings, including elsewhere in the University, in other libraries, in museums or conference centres.

1.2 General Methodology

In order to achieve the main goal of building a model that can detect whether a user is settled or not, we follow the methodology proposed by Mondal [8]. He proposes a general standardised seven-step process for constructing predictive models, in an effort to reduce development time and maintain quality standards. The steps can be summarised as follows:

1. **Business Objective:** Outline what problems the model will solve and what questions it will need to give answers for (Chapter 1).
2. **Goal:** Based on the business objective clearly define the goal of the model (Chapter 1).
3. **Data Selection:** Select and collect data. The effectiveness of the resulting model depends heavily on the quality of the data (Chapter 3).
4. **Data Cleaning & Preparation:** Clean up the data, link the different datasets, validate their correctness and explore the characteristics of the data (Chapters 4 and 5).
5. **Variable analysis and transformation:** Apply transformations, handle outliers and missing data, and apply dimensionality reduction if necessary (Chapter 5).
6. **Model selection and development:** One or more modeling techniques are chosen based on the goals defined in step 2. After ensuring that the assumptions of the chosen techniques hold for the dataset, develop and train the model using the training set. A validation set and an error function can be used to fine-tune the hyperparameters of the chosen models (Chapter 6).
7. **Validation and Testing:** Make predictions using the testing set and calculate the performance of the models using metrics such as accuracy, F1 score etc. (Chapter 7).

1.3 Contributions

The main contributions of the project can be summarised as follows

- The available data are summarised and their main characteristics are outlined, providing insights into properties such as level of noise, variability, etc.
- The risk of revealing identifiable information from the data is assessed.
- A preprocessing pipeline for the data is established, which removes outliers, cleans-up noise, segments the data to sessions and detects the primary device for each user.
- A model that can predict whether a user is settled somewhere in the Library is constructed, after several candidate models were compared. The features for this model are calculated using a window that changes its size depending on the volume of events.
- The model is evaluated by calculating performance metrics on a portion of a separate testing set that has been manually labelled.
- Additionally, as further evaluation, a second model is constructed that uses the output of the settled periods model, to predict whether an access point is located at a study space or not with the results compared against the access points with known locations. Since we expect students to be settled for longer and more students to settle at study spaces than non-study spaces, if this second model has a good result, it is a further indication that the output of the settled periods model is sensible.

Chapter 2

Background

In this chapter we outline what data is available from the University of Edinburgh's Main Library (Section 2.1) and describe what work has already been carried out in the Library for occupancy monitoring using these data sources (Section 2.2). In addition, we give some background on Machine Learning techniques used in this project (Section 2.3) and discuss some relevant work from the literature in crowd monitoring, occupancy tracking and positioning using Wi-Fi data (Section 2.4).

2.1 Data Sources

As part of its effort to improve provision of information to students and staff the University's Main Library has been collecting data useful for occupancy monitoring from three main sources. These are:

- Wi-Fi Wireless Access Points (Section 2.1.1)
- Turnstiles at main entrances and exits of the library (Section 2.1.2)
- Study Space Booking System (Section 2.1.3)

2.1.1 Wi-Fi Wireless Access Points

There are multiple Wireless Access Points (WAPs) spread across each of the floors of the Library building. WAPs are networking hardware devices that allow other devices to connect wirelessly to the network. Most users of the library connect to the library's network primarily to access the internet, making WAPs a potentially effective tool for tracking occupancy. More generally, Ouf et al. [9] suggest that Wi-Fi technologies can be used in place of costly dedicated sensors, such as carbon dioxide sensors, to reliably monitor occupancy at high levels of accuracy.

All device connections and disconnections are recorded for each individual WAP. Each access point covers an area roughly equal to $10m^2$ [1], allowing us to locate an individ-

[1] Communicated to us verbally by Simon Chapple, Head of Data Technology at the University of Edinburgh Information Services.

ual at that level of granularity as they move around the building. It is worth noting, however, that the data can be noisy, for instance including connections to a WAP in a different floor or from users outside the building, if either happened to be in range.

Each connection or disconnection is logged as an event in a table, henceforth referred to as the '*Wireless Access Point log*'. The log keeps track of whether the event represents a new connection (*Start*), a connection with a different Access Point (*Update*) or a disconnection (*Stop*). Anonymised user and device identifiers and an Access Point identifier are also recorded. Each Access Point ID is associated with a physical location within the building. The exact date when the data was collected is unknown, as a further layer of protection of the users' privacy (see Section 3.3).

2.1.2 Entrance Turnstiles

All public-facing entrances and exits of the building are controlled by card-operated turnstiles. Students and staff are required to swipe their university-issued card when entering or leaving the building. For each card swipe, the timestamp, an anonymous user identifier, the location of the turnstile and whether the user entered or exited are recorded. If anonymous user identifiers match the ones that appear in the Wi-Fi log, the two datasets can be linked.

The turnstile data is a reliable indicator of when exactly the user entered or left the building. Using the turnstile data in tandem with the Wi-Fi data allows us to filter out users recorded in the Wi-Fi log while being outside of the building.

2.1.3 Study space Booking System

In order to ensure physical distancing during the Coronavirus pandemic, the library has introduced a new booking system for study spaces. Users have to book a study space in advance, check-in by scanning a QR code when they arrive, and check-out when they leave. The bookings, check-ins and check-outs are all recorded anonymously.

2.2 Current occupancy monitoring methods

The library has already experimented with a few different occupancy monitoring methods, using a variety of data sources. This has resulted in the following occupancy level indicators:

1. **Traffic light system:** A single metric describing how busy the building as a whole is based on turnstile data [15]. An example of how this is shown is given in Figure 2.1.
2. **Floor occupancy indicators:** One value for each floor, representing how busy the study spaces on that floor are, based on the Wi-Fi log [16].
3. **Temperature & Movement sensors:** Sensors attached under each study desk in the lower ground floor of the library building determine whether the desk is

occupied or not. The number of available study spaces in the lower ground floor can be measured and is displayed prominently within the library [17].

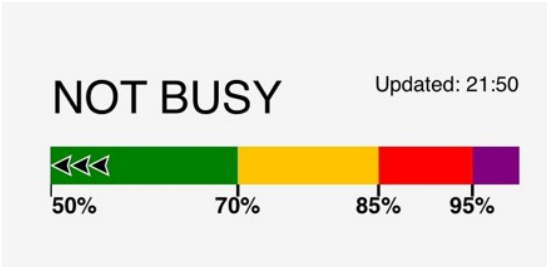

Figure 2.1: The traffic light system, as displayed on the library's website

While the Library has received very positive feedback by students on the information provided by the occupancy monitoring methods mentioned above, the information these metrics provide is not very detailed. They indicate how busy the library overall is or provide an estimate on how busy each floor is. They do not give information to the students such as where specifically they need to go to find a study space more quickly, or how long it can take them to find a study space at any given time. This project aims to work towards this direction, to enable the provision of more precise information both to students and to the Library management. The sensors, while more precise, are costly and require installation of hardware at a large scale. In our project we aim to utilise existing data sources.

2.3 Machine Learning

This section gives a brief overview for the key machine learning methods that are used in this project. In Chapter 6 four candidate models are compared. Their performance is calculated using Stratified k-Fold Cross-Validation (Section 2.3.1). The best model is selected which is a Random Forest Classifier (Section 2.3.2). Logistic Regression is used for building a simple classifier as part of the evaluation of the previous model (Section 2.3.1).

2.3.1 Logistic Regression

Logistic Regression is a common linear classification algorithm. It is common in binary classification problems where an example either belongs in class 0 or 1. It calculates the probability of an example being in class 1 by calculating a weighted sum of the features and then applying the logistic function on it. Therefore, probability of a data point $\mathbf{x}$ being in class 1 is given by:

$$P(y = 1|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x})$$

where $\mathbf{w}$ is the weight vector and $\sigma(\cdot)$ is the Logistic Function $\sigma(t) = \frac{1}{1+\exp(-t)}$. Then, the class y is assigned as follows:

$$y = \begin{cases} 0 & \text{if } P(y = 1|\mathbf{x}) < 0.5 \\ 1 & \text{if } P(y = 1|\mathbf{x}) \geq 0.5 \end{cases}$$

The weight **w** is calculated using maximum likelihood estimation. In other words, **w** is calculated such that the training data (with their given class labels) are most probable.

2.3.2 Random Forest

The Random Forest Classifier is an ensemble classification method. Ensemble classification methods combine multiple classifiers into one model, and the predictions of the individual classifiers are combined using simple majority voting. In the Random Forest, these classifiers are decision trees. Each of these decision trees is constructed using a random subset of the features on a sub-sample of the training set.

According to Ho [14], random forests address the limitation of decision trees that they can easily overfit on the training data, as each individual decision tree is trained on a different random subspace of the feature space.

2.3.3 Stratified k-Fold Cross-validation

A small portion of the training data is usually held out as a validation set, which can be used to adjust the hyperparameters of the model. The model is trained on the training set and the performance on the validation set is calcualated. The hyperparameters that give the best performance on the validation set are chosen.

Géron [5] argues that this approach can sometimes be problematic. If the validation set is too small, the performance metrics for the candidate models could be imprecise. If the validation set is too large, the candidate models are trained on a dataset much smaller than the final model.

A solution to this is k-Fold Cross-validation. The training data is split as evenly as possible into k folds. A fold is simply a subset of the training set. Then k experiments are run. In each experiment one of the k folds is used as the validation set and the remaining $k-1$ folds are used for training the model, as shown in Figure 2.2 for $k = 5$. Performance metrics are calculated for each experiment and then averaged across all experiments.

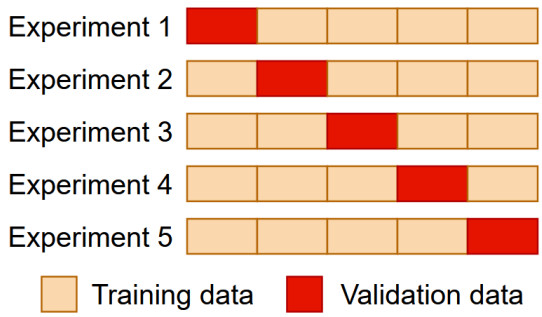

Figure 2.2: 5-fold Cross-validation

Cross-validation is Stratified when the distribution of the target class in each fold is as similar as possible.

2.4 Related Work

The significant uptake of Wi-Fi enabled devices during the last two decades has presented opportunities for occupancy monitoring of buildings, crowd management at large-scale events and indoor localisation for enhancing the provision of services. These are areas that have been widely explored in the literature.

Di Flora and Hermersdorf [3] proposed a scalable solution for positioning users in indoor environments using a novel technique based on received signal strength. This enables providing the user with location-based services such as wayfinding in airports or interactive activities in museums. Krumm and Horvitz [6] also developed a system based on received signal strength to determine whether a user is still or moving and used a Hidden Markov Model (HMM) to infer the user's location.

Savnik [11] developed a position tracking system by collecting Wi-Fi probe request frames from strategically positioned receiver stations with the aim of using the results to inform allocation of resources in a 'Smart City'. He experimented with two position estimating approaches: proximity based positioning and trilateration. A similar approach was employed by Chilipirea et al. [2] who also tested their approach during a large-scale festival and found it to be prone to erroneous readings. They proposed various techinques to mitigate this and clean up the raw detections to enable their use for crowd analytics.

Mohottige et al. [7] explored the feasibility of using metadata from Wi-Fi access points to estimate occupancy levels at room-level in a university campus. They compared the performance of their Wi-Fi based occupancy estimation to a hardware-based approach, using Beam counters, and concluded that they provided comparable results. The low installation and maintenance costs of the Wi-Fi approach made it a more attractive option.

The examples above demonstrate the increased interest in the literature for using Wi-Fi data for the purposes of crowd management and occupancy monitoring. However, different types of data can be collected from Wi-Fi access points. The examples above make extensive use of signal strength in their implementations. The data available (see Chapter 4) for our project are much simpler - they only include which access point each device is connected to, which results in a different approach from the above being pursued. This is both a challenge, as well as an opportunity to adapt our findings to different settings.

Chapter 3

Access to Data

The data for the project is provided by the University of Edinburgh IoT & Innovation Service and the University of Edinburgh Library & University Collections. The Library has started collecting data from Wireless Access Points and Entrance Turnstiles since May 2019 and this is the first time data was shared for a project outside the Library itself. The collection and use of this data is governed by strict privacy policies that have been discussed and agreed with the student body through a series of consultations and focus groups. Maintaining the privacy of the students and staff entering the library is paramount to allowing its safe and free use. Therefore, it was considered essential to set a good precedent and conduct the data sharing in a formal way with users' privacy at the forefront.

Firstly, we prepared a Data Request Document in an effort to clearly document the data needs and formalise the relationship with the data providers (Section 3.1). Due to the sensitive nature of our data and our initial plans for recruiting human participants (Section 3.2), we followed the formal Ethics Approval process of the School of Informatics (Section 3.3). We discuss the details of the process to obtain the data in the following sections.

3.1 Data Request

In the Data Request (available in Appendix A) we outlined what data we need, the amount of data that is requested (how many days worth of data), what the purpose of using the data is and how we are going to safeguard users' privacy.

The data that we requested is:

- Wireless Access Point (WAP) logs
- Turnstile data for the main entrance of the Main Library
- Floorplans and metadata for the locations of Wireless Access Points

All data is de-identified with all User IDs and Device MAC addresses removed and replaced by a simple, randomised numeric IDs that have no association with the users.

In addition, it is not made known to us when the data we are given access to was recorded. Only the time of the day for each recorded event is known. A week's worth of data for the Wi-Fi log and the Turnstile log was requested initially.

A three-person board representing the data providers was set up to review and approve the Data Request and oversee any ethical issues that may arise during the project. The board consisted of:

- Kirsty Lingstadt (Head of Digital Library at the Library & University Collections)
- James Stewart (Lecturer at the School of Social and Political Science & Leader of the Ethics and Governance group for the IoT Research and Innovation Service)
- Simon Chapple (Head of Data Technology at the Information Services Group & IoT Research and Innovation Service)

In order to ensure that the project was conducted in an ethical and privacy-oriented way it was agreed that the project would undergo the Informatics Ethics Approval process and that a preliminary data analysis would be undertaken to ensure that individuals with consistent movement patterns (e.g. Library staff) cannot be identified.

3.2 Human Participants

Initially, there were plans to recruit participants in order to get their consent to have the de-anonymised data corresponding to them obtained. Then, they would be interviewed or asked to complete a survey so the data corresponding to them could be matched to their actual activity in the main library which would assist in the evaluation of our methodology and in getting further insights on what the anonymised data can offer.

For this reason a Participant Information Sheet (PIS) was prepared. The PIS gives general information about the purpose of the project, what data will be collected, the reasons why human participants are being recruited and how their data would be stored. Any collected data will be stored on storage solutions that are provided by the School of Informatics or the University for this purpose.

It was not possible to recruit human participants this year as originally planned, due to restrictions related to the ongoing COVID-19 pandemic, which did not permit physical access to the library for non-essential purposes. Therefore, we had to adapt and only use the anonymised data. Some data was labelled manually using our intuition on how people move around and information on the layout of the library. Now, recruitment of human participants is planned to take place as part of future work.

3.3 Ethics Approval

Because of the sensitive nature of the data and the initial plan to recruit human participants (Section 3.2), it was necessary to undergo the Informatics Ethics Approval pro-

cess. In the ethics application we outlined what data the project involves, and how it will be used. We explained how privacy is safeguarded by having all personal identifying information removed before obtaining access and by using data without knowledge of when it was collected.

As part of the process a data management plan was prepared to specify what data is used. It outlined that the data will be stored on an encrypted file system and will be deleted after 2 years, which is how long this project is expected to last. A record of where the data has been stored is kept to ensure that it can be safely deleted after the completion of the project. It also explained how access to the data was arranged and how any ethical issues are managed as described in Section 3.1. Additionally, the Participant Information Sheet which was prepared as part of the initial plans to recruit human participants was submitted for approval. Our ethics application was approved with reference number 2019/66756.

3.4 Timeline of access

The Data Request was submitted in October and the first sample of the data was received in December. The first sample (Dataset A) includes the Wireless Access Point Log for the Main Library for some unknown 24-hour period. The data has been de-identified in advance before being handed over to us. The second sample (Dataset B) includes the Wireless Access Point log, in a format identical to the first sample but for a different unknown 24-hour period. User and Device IDs do not correspond to the IDs from the first sample. In other words, if a user ID appears in both datasets, this does not imply that the corresponding entries refer to the same user.

A summary of when data was received is shown in Table 3.1. Initially, there were also plans to receive a final sample of data, from 7 consecutive days which would include a field indicating whether each event was recorded by a student or staff member. As the approval to get access takes time and due to high workload in the Library, given the difficult circumstances during the current pandemic, we were not given access to this final sample in time to be included in the project.

Date	Description
12 Oct 2020	First version of Data Request submitted
27 Oct 2020	Data Request resubmitted after some minor amendments
04 Dec 2020	First Sample of the Wi-Fi log for the main library **(Dataset A)**
11 Jan 2021	Floorplans with locations of some of the Access Points
8 Feb 2021	Second Sample of the Wi-Fi log for the main library and the relevant turnstile data for some of the users in the Wi-Fi log **(Dataset B)**

Table 3.1: Timeline of access to data

Chapter 4

Data Overview

In this chapter we outline the attributes of the available data (Section 4.1) and explore some of the main characteristics of the data (Section 4.2) which will influence our work on preprocessing and building the classification model. Finally, some analysis is carried out to explore the possibility of extracting infomration that could lead to a user being identified from the anonymised data (Section 4.3).

4.1 Description of Data

We describe the attributes of the available data which include the Wireless Access Point Log (Section 4.1.1), the turnstile log from the entrances and exits of the Library (Section 4.1.2) and information on the locations of access points in the buildings (Section 4.1.3). This data will then be used for some initial exploratory analysis. Most of the initial exploratory analysis is presented using Dataset B. The Dataset A, which is set aside to be used for evaluation, has very similar characteristics.

4.1.1 Wireless Access Point Log

The Wireless Access Point log (WAP log) consists of a sequence of events accross multiple users and devices. Each event has the following attributes:

- **Timestamp:** The time when the event was recorded. It is given as Unix time. The date corresponding to the timestamps is artificial and does not represent the date when the events were recorded.
- **UID:** A numeric user identifier. It can be used to identify individual users in the dataset, but has no relationship to the real user identifier.
- **Device:** A numeric device identifier. It has no relationship to the MAC address or any other real identifier of the device. It is only for the purpose of distinguishing different devices.
- **Event Type:** Whether the event is a 'Start', 'Update' or 'Stop' event. More details about the types of events are given in Section 4.2.3.

- **Reason:** This field is only populated when the event is a 'Stop' event. It indicates whether the connection was terminated due to a user request or an idle timeout.
- **Network:** Indicates to which network the device was connected. This can be either 'central' or 'eduroam', the Service Set Identifiers (SSIDs) of the two main networks of the library.
- **Wapmac:** The physical MAC address of the Wireless Access Point the device is connected to.
- **Location:** A description of the location where the WAP is located. This could be either a floor (e.g. 'Main Library 5th floor') or a room (e.g. 'Main Library Room 1.27').

4.1.2 Turnstile Log

The turnstile log is simpler compared to the WAP log and only contains 3 fields for each event:

- **Timestamp:** Similar to the timestamp for the WAP log. The time when the event was recorded given as Unix time. As for the WAP log, the date corresponding to the timestamps is artificial and is not representative of the date the events were recorded.
- **UID:** A user identifier for the user the event was recorded for. It is not associated with any real user identifiers and is only used for distinguishing between different users. For the same user, it is the same UID that is recorded in the WAP log.
- **Turnstile:** The location of the turnstile where the event was recorded. These can either be the Main Entrance or Exit, or the turnstiles between the Cafe and the Library.

The turnstile log was only available for Dataset B. Even then, it was not possible to match all users from the Wi-Fi log with users in the Turnstile log. Out of the 7697 users in the WAP log, there are entries in the Turnstile log for 4303 of them. After enquiring with the data provider we found out that when they store the data, the UIDs are not linked between the Wi-Fi and turnstile data. They have a way to link them, but it was not successful for all users. The number of devices, users and events in the WAP and turnstile logs for both datasets A and B are summarised in Table 4.1.

4.1.3 Floorplans & location metadata

We were given access to a set of floorplans showing the precise locations of some of the access points, as well as to metadata mapping the MAC addresses of access points, which are used in the WAP log, to the WAP IDs. The first character in a WAP ID designates the floor where it is located. For example, 'G22' is on the ground floor and '411' is on the 4th floor.

	Dataset A	Dataset B
Events in WAP log	229,909	309,505
Devices in WAP log	6,995	12,197
Users in WAP log	4,862	7,639
Users in turnstile log	0	6,231
Users in both	0	4,303

Table 4.1: Summary of the number of events, users and devices in the WAP and turnstile logs for Datasets A and B. For Dataset A only the WAP log is available.

Out of the 217 access points only the location of 107 was given from the available floorplans. The locations of some access points, such as the ones on the ground floor, were known at room-level. Given that in the ground floor access points appear to be placed in a way such that adjacent access points had consecutive IDs it was possible to identify the locations of a further 16 access points in the ground floor. The locations of Access Points in the Ground Floor is shown in Figure 4.1. In red are the Access Points their location was known and in blue are the access points the location of which was estimated. Pandemic restrictions prevented us from physically confirming the locations of the APs.

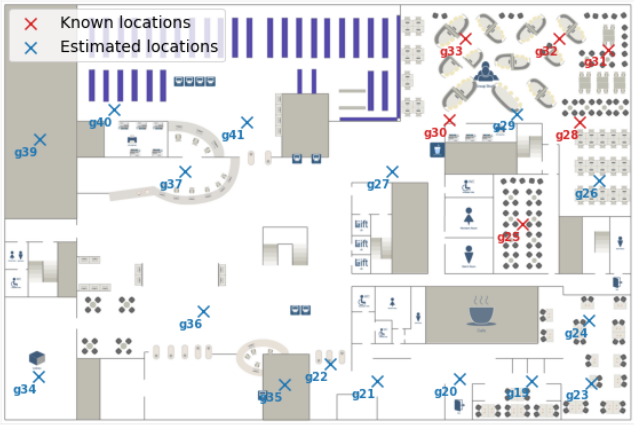

Figure 4.1: Locations of Access Points on the Ground Floor

4.2 Main Characteristics of Data

In this section, we explore some key characteristics of the data that will inform the next preprocessing steps, using Dataset B which is going to be used for training the model. Dataset A, which has been set aside for evaluation has almost identical characteristics as Dataset B, unless stated otherwise. Some of the characteristics we will explore are the number of devices per user and a first estimate on the amount of time they are connected to the network (Section 4.2.1), which will inform our approach for outlier removal and the need to identify one of the devices as 'primary' to be used for classi-

fication. We also define what we mean by 'event duration' (which is different than the Duration field already included in the log) in Section 4.2.2 and comment on its size. In Section 4.2.3, the three types of events are described and we explore the possible event type sequences that can appear, which is used later in Section 5.1 for segmenting the WAP log into sessions. Finally, in Section 4.2.4 we discuss the common access point transitions, by performing N-gram analysis and discuss what that could mean in terms of noise.

4.2.1 Devices and Users

During the 24 hours covered by Dataset B, 7679 unique users connected to the Main Library's network across 12197 devices. Most users, (52%) connect with two devices to the network, whereas 45% connect with just one device. A further 4% connected with three devices, while the number of users with more than 3 devices is negligible (4 users in total). Interestingly, in Dataset A 59% of the users connect with just one device, which is significantly higher compared to Dataset B. This could indicate that the datasets were recorded in different times of the year (e.g. It might be more likely for users to bring their laptop during the exam period).

A significant number of users appear to have connected for a very short period of time, as low as 5 minutes, based on the difference between their latest and earliest timestamps as shown in Figure 4.2. These could be users walking outside the building who only connect for a few minutes while coming in the range of an access point. Therefore, the actual number of users who are using the Library is likely to be smaller.

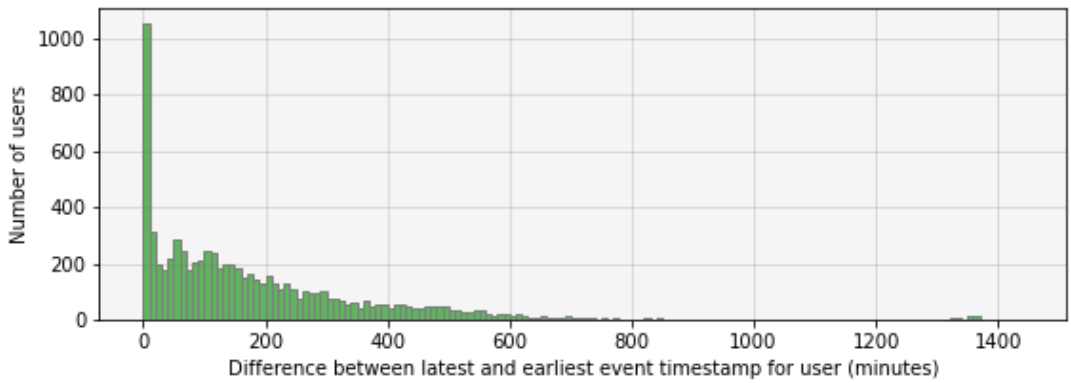

Figure 4.2: Timestamp Range

4.2.2 Event Duration

The recorded events are certain instants in time. Therefore, by 'Event Duration' we refer to the time from the current event until the next event is recorded (This is different from the 'Duration' field in the WAP log which measures the time since the last *Start* event - see Section 4.1.1). Hence, it is the time period for which the device was connected at the particular AP recorded in the event, if the WAP log for a particular

device is split into sessions where the device is known to be within the Main Library's network.

A large proportion of events have very short event duration as shown in Figure 4.3. More specifically, 44% of events have duration of less than a minute. This could indicate that a large number of events is recorded in a short period of time when someone is moving in the library.

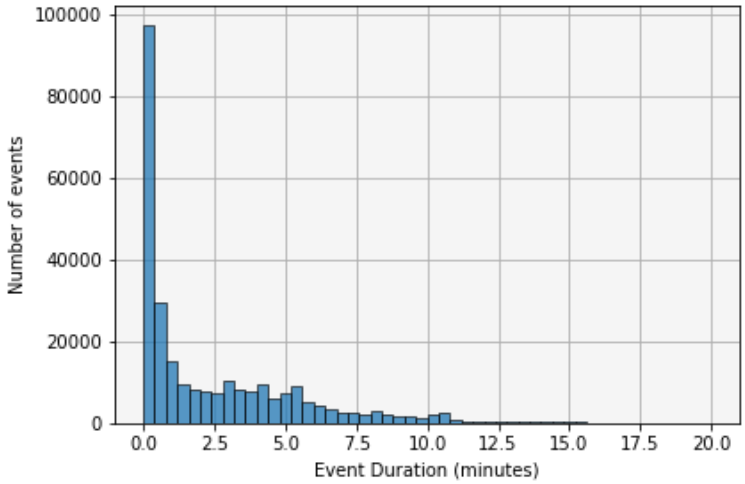

Figure 4.3: The event duration for all events recorded within the 24-hour period (Event durations longer than 20 minutes are omitted which account for 6% of all events)

4.2.3 Types of Events

There are three types of events in the Wi-Fi log: *Start*, *Stop* and *Update*. The majority of events are *Update* events with 82% of the total number of events with the rest roughly equally split to *Start* and *Stop* events. A *Start* event indicates that the device has started a new connection to the university's network. A *Stop* event indicates that the connections is terminated. An *Update* event is recorded when the device moves from one access point to another. Sometimes, an *Update* event is recorded even when the device does not move to a different access point. This is referred to as a 'heart-beat' event which indicates that the connection is still active. It is not clear how often these 'heart-beat' events are recorded and for many devices they are not recorded at all.

One would expect that the first event recorded in the log for a particular device would be a *Start* event and the last a *Stop* event. However, this is not always the case. Only 59% of the first events recorded by the devices are *Start* events and 43% of the last events recorded by the devices are *Stop* events. In addition, a large proportion of *Start* events were preceded by an earlier event a few minutes before as shown in Figure 4.4. Similarly, for a large proportion of *Stop* events, another event follows after very few minutes as shown in Figure 4.5. In both figures, the orange bar represents the total number of events such that the time difference was at least 30 minutes. In the former plot this accounts for just 9% events and for the latter for 10% of events.

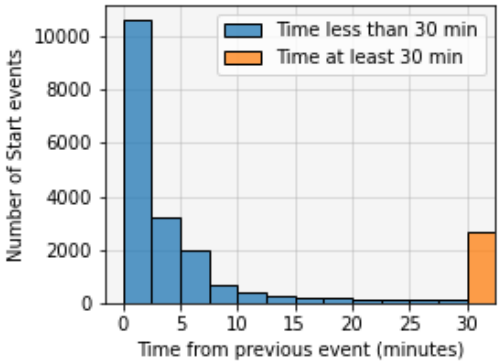

Figure 4.4: Time between *Start* events and the event preceding them

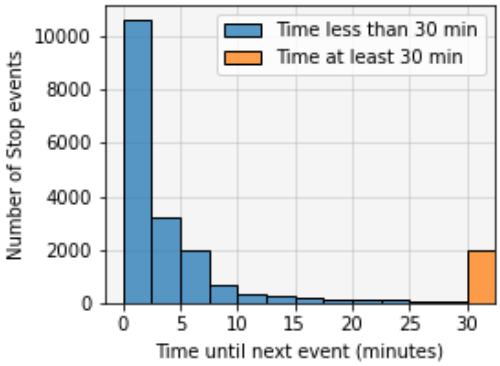

Figure 4.5: Time between *Stop* events and the event after them

These observations could suggest that for a significant proportion of devices, the first event recorded when they connect to the Library's network might not be a *Start* event and the last might not be a *Stop* event. This is explained by the fact that the available WAP log is a portion of the WAP log for the entire University network corresponding the Main Library's WAPs. It is likely that devices are already connected to the University's network when entering the library, hence their first event would be an *Update*. Therefore, the event type is not a reliable indicator for when a user entered or left the Main Library.

4.2.4 Access Point Transitions

We can discover which access points sequences are most common by treating the event logs for each user as a sequence of access points and then detecting the most common N-grams. N-grams are common in computational linguistics and are a contiguous sequence of N, in our case, access points.

The four most common access point trigrams ($N = 3$) and tetragrams ($N = 4$) from our event log are shown in Table 4.2. Some of these appear to be cycles where the device connects to a different access point and then reconnects back to the previous one. In addition, 20% of all trigram occurrences and 9% of all tetragram occurrences involve only one pair of access points. This could suggest that even when a user is stationary their device can jump between different access points that are nearby, which in turn results in new events being recorded in the log.

4.3 Possibility of user identification

During our discussions with the data provider, the possibility of identifying individuals with unique activity patterns was raised. This could be a concern especially for staff members who have a more distinctive usage compared to students. They tend to have access to areas not accessible by the public and have set working hours. A typical working day lasts from 9 in the morning to 5 in the afternoon.

Most common trigrams		Most common tetragrams	
Trigram	Frequency	Tetragram	Frequency
145, 144, 145	440	145, 144, 145, 144	314
144, 145, 144	435	144, 145, 144, 145	291
G36, G37, G36	425	427, 432, 427, 432	120
G36, G37, G35	410	432, 427, 432, 427	114

Table 4.2: The four most common access point trigrams and tetragrams shown as sequences of access point IDs

In this section we examine one such example of a user in an attempt to establish to what extent they can be identified by the available data.

We consider one access point (AP with ID G39) which is in a part of the library not accessible by the public. Despite this, we expect devices from the area surrounding the room to still be able to connect to the access point. Figure 4.6 shows the total time each device spent connected to that particular access point.

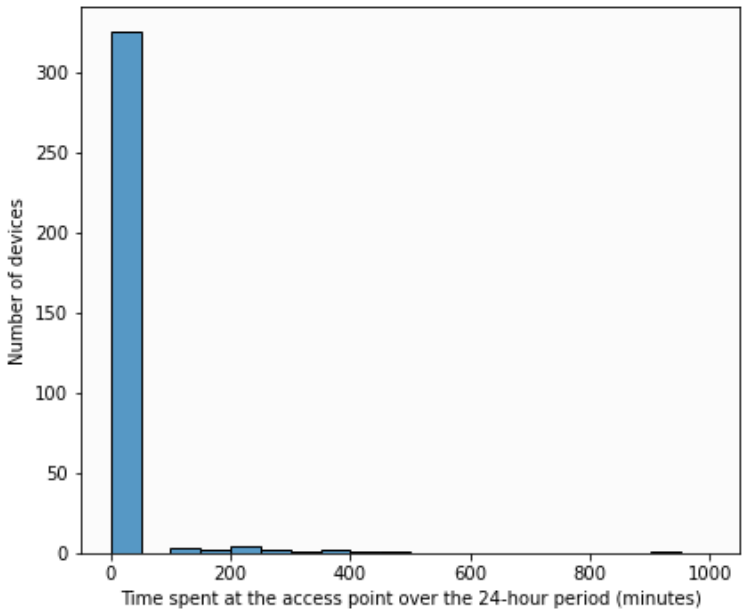

Figure 4.6: Total time devices were connected to a particular access point (G39) located in an area of the library not accessible by the public during the 24-hour period

The vast majority of users were connected to the access point for less than 25 minutes in total during the 24-hour period, which is what we expect from users moving in the surrounding area. There is a single device, however, that was connected for approximately 15 hours. By examining the WAP log corresponding to it, it appears to be a stationary device that is connected continuously from midnight until 5 PM. Then, there is a cluster of devices that were connected for 100-500 minutes in total. Figure 4.7 shows the access points visited by one of those devices which was connected to the

access point for 445 minutes (7.5 hours) in total. Figure 4.8 shows the path followed by the same device during the day on a floorplan of the Ground Floor.

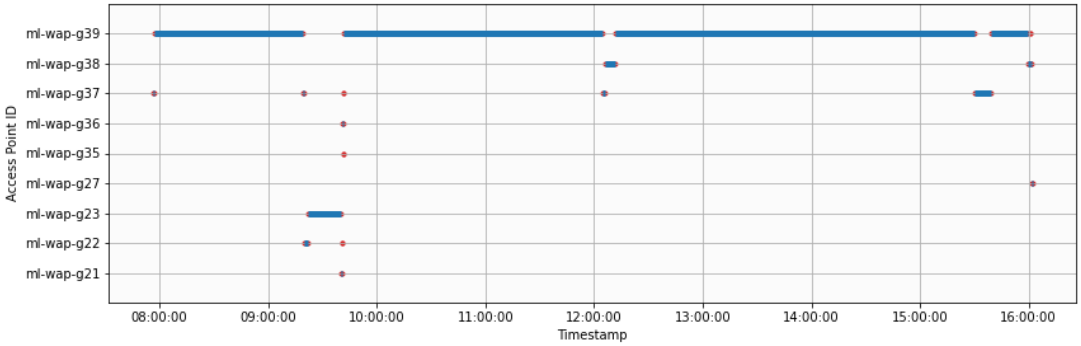

Figure 4.7: The access points visited by the device connected for 7.5 hours in total to access point G39

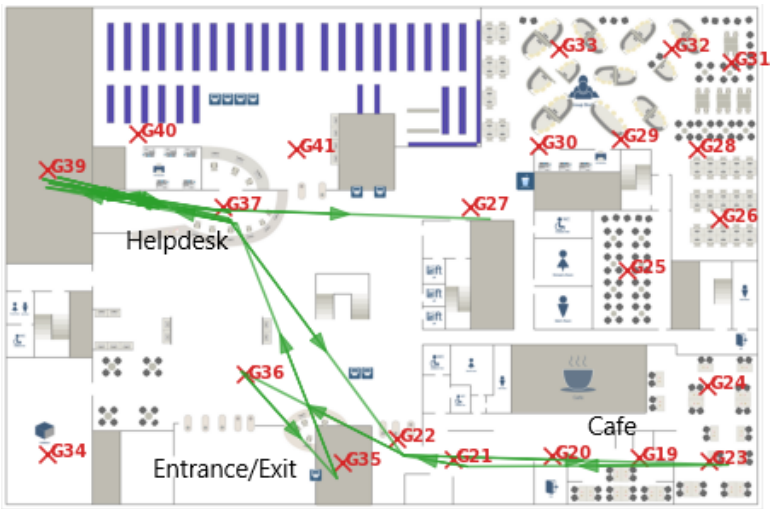

Figure 4.8: The path on the ground floor followed by the device connected for 7.5 hours in total to access point G39

According to this, the user arrived at 8 AM and left at 4 PM (i.e. 9 AM to 5 PM in British Summer Time) which is consistent with a working day. The user spent most of their time at the access point G39 which is located in a room next to the Helpdesk (see Figure 4.8) while also visiting G37 on two occasions which is the Helpdesk. They also visited the cafe (G23) and remained there for about 20 minutes.

These observations suggest that this user is a member of staff, possibly working at the helpdesk. Despite the fact that this user has unique movement patterns compared to other users, without any additional information and due to lack of knowledge of when the data was recorded, it is not possible to further identify the user, even if the list of members of staff manning the helpdesk was known. In addition, since data is available only for a single day and user IDs in Dataset A do not correspond to user IDs in Dataset B, it is impossible to observe any daily patterns. Therefore, with the available data it is very unlikely to extract any further identifiable information about the user's identity.

Chapter 5

Data Preprocessing

Based on our observations on the Wireless Access Point log dataset outlined in Section 4.2, we have established a pre-processing pipeline to clean the data and extract useful, higher-level information. This is an essential first step to support the model that detects when a user is settled at a particular area in the library, which we go on to describe in Chapter 6. In this pre-processing stage, the WAP log for each device is first split into 'sessions' (Section 5.1), which enables us to treat each visit of a user to the library separately, by detecting when the user is actually in the building. Sessions that are considered outliers (Section 5.5) are removed, as our focus is on users who find a study space in the library and remain there for a significant amount of time. Consecutive events are merged where possible (Section 5.2) and events that are considered noise are removed (Section 5.3). In addition, a method of picking one of the devices for each user as the 'primary' device is described (Section 5.4). The WAP log corresponding to this 'primary' device is the one that then used for predicting the settled periods.

The performance of any model depends heavily on the way the data is preprocessed. Our model that detects settled periods which is described in Chapter 6 uses a window, with size that changes adaptively based on the volume of events. Therefore, it was important to remove any unnecessary events by merging 'heart-beat' events (Section 5.2) and removing some of the noise (Section 5.3).

5.1 Session splitting

As discussed in Section 4.2.3, event types (*Start*, *Update*, *Stop*) are not a good indication of when a user entered or left the library building. Therefore, it is necessary to develop a way to split the WAP log for a particular device into several 'sessions'. Sessions are subsets of the entire WAP log corresponding to a particular device, such that they describe a singular visit to the library from the device's user. For example, a user can visit the library in the morning for two hours, then leave and return to the library again in the afternoon for a further three hours. These two visits should be treated as two separate sessions.

In order to divide into sessions, we split the log when it appears that the device has connected to an external access point (i.e. one that does not belong to the main library network). This can be detected through *inconsistent event sequences* in the device's WAP log (also see Section 4.2.3). For instance, Start and Update events cannot be followed by a Start event and Stop events are normally only followed by a Start event. If this is not the case, it indicates that we are missing some events that occurred when the device was connected to a part of the university network outside the library. More specifically, the possible inconsistent event sequences are the following:

- Start event preceded by a Start or Update event (missing Stop event).
- Update event preceded by a Stop event (missing Start event).
- Stop event preceded by a Stop event (missing Start event).

In such cases, we decide to split the session at the point between the 2 events in the inconsistent sequence. Table 5.1 shows an example of an event log split to sessions based on an Update-Start sequence. An event was recorded between events 2 and 3 outside the main library. Therefore, a new session starts at event 3.

#	Timestamp	Event Type	WAP ID
1	11:54:22	Update	153
2	13:02:38	Update	156
3	15:00:02	Start	306
4	15:00:40	Update	316

Table 5.1: Example of a session split based on an inconsistent event sequence (*Update - Start*) shown by the red line

One other way to establish whether an event was recorded outside the library is using the *Duration* field (see Section 4.1.1). The Duration field measures the time since the most recent Start event. If the difference between the Duration of the previous event and the Duration of the current event is not equal to the difference of their timestamps, this means that there is a missing Start event that occurred between these two events. Therefore, the current event is the start of a new session. One example of this is shown in Table 5.2 where event 3 is the start of a new session.

Although these conditions give us a reasonable split of user sessions, observing some examples from the data it is apparent that WAP logs for some devices can be split to even further sessions. This is particularly obvious in cases of consecutive events with a large difference between their timestamps. If the previous event was a Stop event and the current event was a Start event, neither of the two criteria above can apply. We cannot always split on a Stop-Start sequence, as the time between adjacent Stop and Start events is less than 5 minutes for 63% of them, which indicates the presence of many brief disconnections. A threshold of 2 hours has been set as the maximum amount of time between two consecutive events, so that they can be considered to belong to the same session. Otherwise, it would mean that the user did not move at all, the device did not connect to a different access point (despite the high density of

#	Timestamp	Event type	WAP ID	Duration (min)	Timestamp diff. (min)	Duration diff. (min)
1	11:44:15	Update	423	25.7	2.8	2.8
2	11:46:51	Update	G37	28.3	2.6	2.6
3	12:00:35	Update	G19	1.1	13.7	-27.2
4	12:02:14	Update	G19	2.8	1.7	1.7

Table 5.2: Example of a session split based on the timestamp difference compared to the previous event not matching the duration difference compared to the previous event shown by the red line

access points), and a 'heart-beat' event was not registered for at least two hours. This threshold was selected empirically, and can be easily adapted in our model upon the opportunity to validate it in practice. In the case where consecutive events are at the exact same access point, hence the second being either a 'heart-beat' Update event or a Stop event, we increase this threshold to 4 hours, since it is more likely that the user may have remained at the same location during this time.

These criteria result in 11,375 sessions from 6,995 devices in Dataset A and 15,108 sessions from 12,197 devices in Dataset B.

5.2 'Heart-beat' event merging

Events are typically recorded when a user moves from one access point to another. However, in the access point log we often encounter consecutive events which are at an identical location. Upon querying this with the data provider, we were informed that sometimes access points record so-called "*heart-beat*" events as a confirmation that the session is still active.

However, these heart-beat events are not recorded consistently and it is not clear after how much time they are recorded. Since they do not provide any useful information after we have split each device's log to sessions, we remove them from the log. This filtering process reduces the size of the Wireless Access Point log reduces the size of the WAP log in Dataset A by 78,303 events and the size of the log in Dataset B by 107,390 events.

5.3 Removing noise

As was observed in Section 4.2.4, transitions between specific pairs of access points are common. Similarly, we have observed a large number of short Stop-Start gaps, as discussed in Section 4.2.3. These could happen due to the user being in close proximity to multiple access points, whose coverage overlaps. As part of the preprocessing pipeline, we remove some of that noise during long time periods of low activity. In other words, in periods where events are sparse and a brief transition to a different

access point or a brief Stop-Start gap is observed, it is removed. More concretely, this is done using the following rules:

- If the duration of an event is less than 5 minutes, the previous and next events were recorded at the same access point and the sum of the durations of the next and previous events is at least 10 minutes, then the event is removed.
- If a Stop event is followed by a Start event within 5 minutes and the durations of the previous event and the Start event are greater than the time between the Stop and the Start events, and both the Stop and the Start events were recorded at the same access point, then both events are removed.

This does not remove all noise that is present in the WAP log. For example, cycles of three access points are not removed. However it manages to remove many instances of the common types of noise described above, reducing the size of the WAP log by roughly 35,000 events. Figure 5.1 shows an example of a session before and after it was preprocessed in the way described in this section.

In Section 6.2 we discuss the use of an adaptive window for the model, with a size that depends on the volume of events. Shorter events have a greater effect on its size, which in turn influences the classification performance. Therefore, with this preprocessing step our primary aim is to remove short events that are considered noise, for which the given thresholds of 5 and 10 minutes are sufficient.

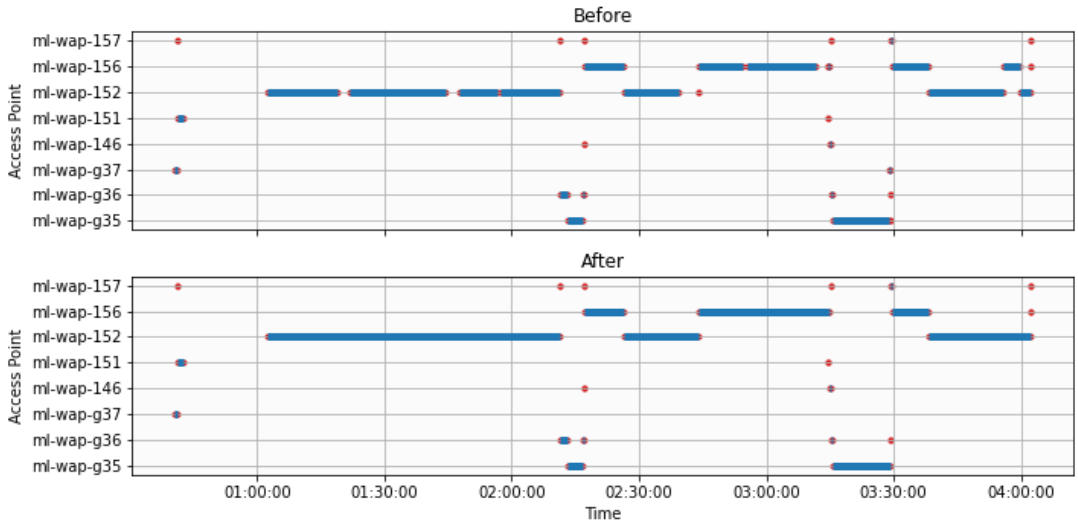

Figure 5.1: An example of a session before and after applying the 'removing noise' preprocessing step

5.4 Determining the primary device

Almost half of the users connect to the network with at least two devices, whilst we have observed users with up to twelve devices. Usually these are at least a mobile device such as a phone or a smartwatch, and a laptop which is usually used on a desk.

We consider the mobile device as 'primary' since it is typically always connected (in contrast to a laptop that connects when it is switched on) and thus from the portion of the WAP log corresponding to it we can get a better idea of the user's movement patterns in the library and extract more useful information, which can then be used as input to our model in Chapter 6.

Figure 5.2 shows an example of the activity for the two devices that belong to a particular user. It is obvious that Device 1 must be a mobile device since it visits significantly more access points across multiple floors. In contrast, Device 2 during the same time period connected only to two access points. In this example, both devices are connected to the network for roughly the same time period (Device 2 was connected for slightly less time). It appears that the time to find a study space and the time to leave the Library is much smaller compared to the time spent settled at a study space.

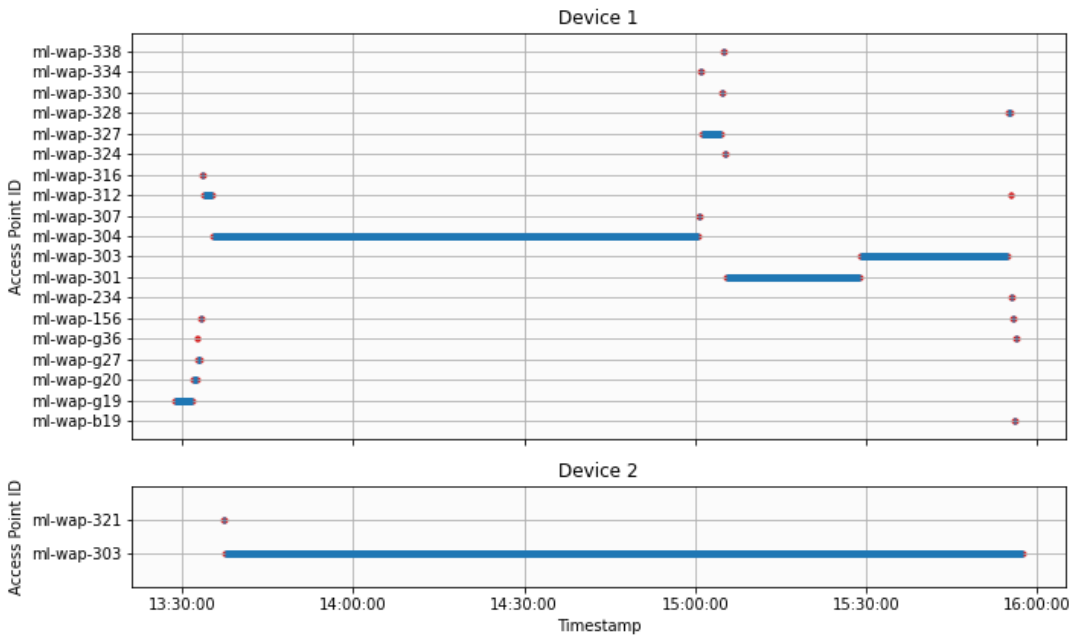

Figure 5.2: The access points visited by the two devices that belong to the same user during a certain time period

In order to confirm whether our observations on the example in Figure 5.2 apply as a general rule, we observe the difference between the number of access points visited by the device which visited the most access points and the device that visited the least access points in the same session. The number of access points visited by the device that visited the last access points for each user is shown as a proportion of the number of access points visited by the device that visited the most access points in Figure 5.3. Figure 5.4 shows the total connection time of the device that was connected for the least amount of time as a proportion of the total connection time of the device that was connected for the most amount of time for the same user. The greater these ratios, the more similar the two values for the two devices are. We want to pick the variable that for the majority of pairs of devices, it is as different as possible. Hence, the ratio would be closer to 0 than 1 for the majority of pairs.

For the majority of users, the first device visits a much larger number of access points compared to their device that visits the fewest access points (Majority of ratios closer to 0 than 1 in Figure 5.3). This is not the case for the time the devices spend connected, as for most users their least and most used devices spend a similar amount of time connected to the library's network (Majority of ratios closer to 1 than 0 in Figure 5.4). Therefore, we treat the device that visited the most access points as the 'primary' device.

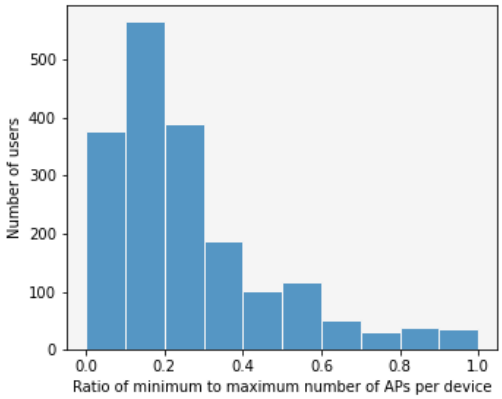

Figure 5.3: The ratio of the number of access points visited by the device that visited the least number of access points, to the number of access points visited by the device that visited the most number of access points for each user with at least two devices

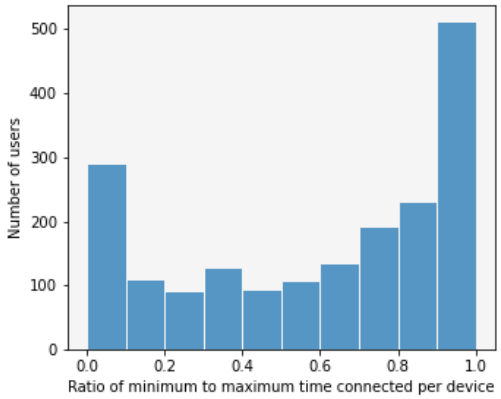

Figure 5.4: The ratio of the total time connected for device that was connected for the least amount of time, to the total time connected for the device that was connected for the largest amount of time for each user with at least two devices

5.5 Outliers

A significant proportion of sessions last only for a few minutes. As shown in Figure 5.5, 11% of sessions are less than 5 minutes long. In addition, 4% of sessions consist of just one event. As an event represents a single time point, it does not provide us with useful information on its own as, for example, it is not possible to extract for how long a user was within range of a particular access point. Therefore, we remove sessions that consist of just one event. We also remove sessions that are less than 5 minutes long, as our focus is on students that spend a considerable amount of time in the library. As we get more evidence, this threshold can be adjusted. Alternatively, if in the future it is possible to link all users in the WAP log to their corresponding entries in the turnstile log, it will be possible to use the turnstile log to filter out users who did not enter the building.

The access points where the events that were discarded as part of outlier sessions (Sessions with one event or less than 5 minutes long) are shown in Figure 5.6. The majority of these events were recorded at access points around the perimeter of the main library in the basement, ground and first floors. More specifically, in access points in the basement facing towards busy pedestrian routes, which are located in the south-west of the

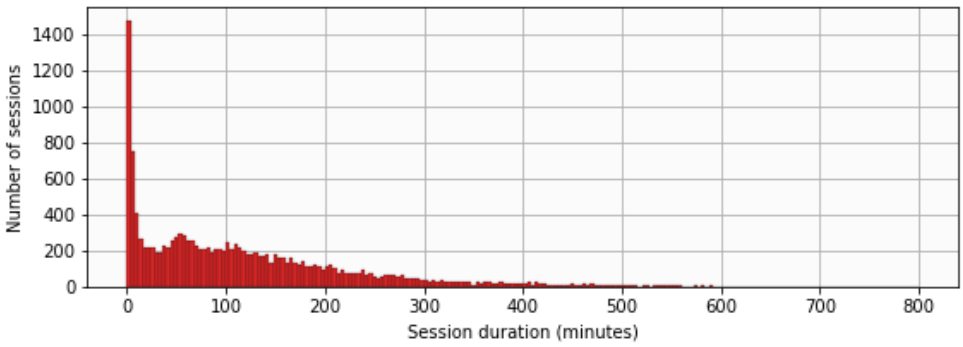

Figure 5.5: The duration of all sessions in minutes

library (top-right of the Figure), or access points on the ground floor and the balcony on the first floor facing towards the north of the building (bottom of the Figure), where there is a busy pedestrian route as well. This suggests that most of these events might be recorded by users walking outside the building, who are not of interest.

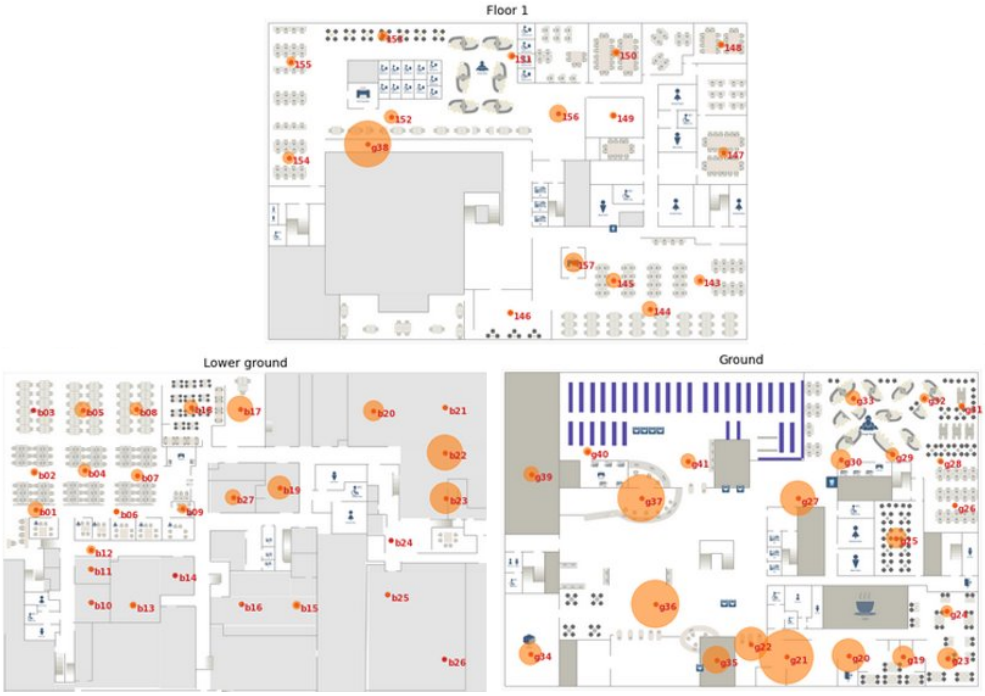

Figure 5.6: The number of events that were removed as outliers per access point represented by the area of the circles. 39 out of the 40 access points with the most removed events were on the Lower Ground, Ground or First floors

Chapter 6

Detecting settled periods

In this chapter, we discuss the construction of a predictive model that can classify events as 'Settled' or 'Not settled'. If an event is classified as 'Settled' this means that the user is estimated to be almost stationary at a particular area of the library between the time the event is recorded and the time the next event is recorded, i.e. during what we consider the 'duration' of the event. First, we define the concept of a '*dominant*' access point (Section 6.1) and discuss the use of an event window which adapts its size depending on the volume of events (Section 6.2), during which several metrics that are used as features for the model are calculated (Section 6.3). Different machine learning algorithms are compared by calculating the performance of each model using Stratified 5-Fold Cross-validation. The dataset has already been preprocessed in the way described in Chapter 5. For each user only the primary device is kept.

6.1 Dominant Access Points

During a session, the device connects with multiple access points. However, for the vast majority of sessions, a significant proportion of the session's duration is spent at a small number of access points. For many sessions, as shown in Figure 6.1, a significant proportion of the duration of the session can be attributed to just one access point. In addition, for virtually all sessions, at least 80% of the session's duration is spent at just three access points. For example, in the session displayed in Figure 6.2, a large proportion of the session's duration is spent at access points 240 and 241.

When a user is settled at a particular area of the library, it is common that their device alternates between a few access points around them, instead of just a single access point as there is a higher density of APs in most study spaces. Hence, it is useful to consider multiple 'dominant' access points for some sessions. As dominant access points we consider the access points where a given device was connected for at least 80% of the session's duration. The number of dominant access points is the smallest needed to reach 80% of the session's duration, which means that the dominant access points are always the access points with the highest share of the session duration. Therefore, if the device was connected to a particular access point for at least 80% of a session's duration, then only that access point is dominant. The number of dominant access

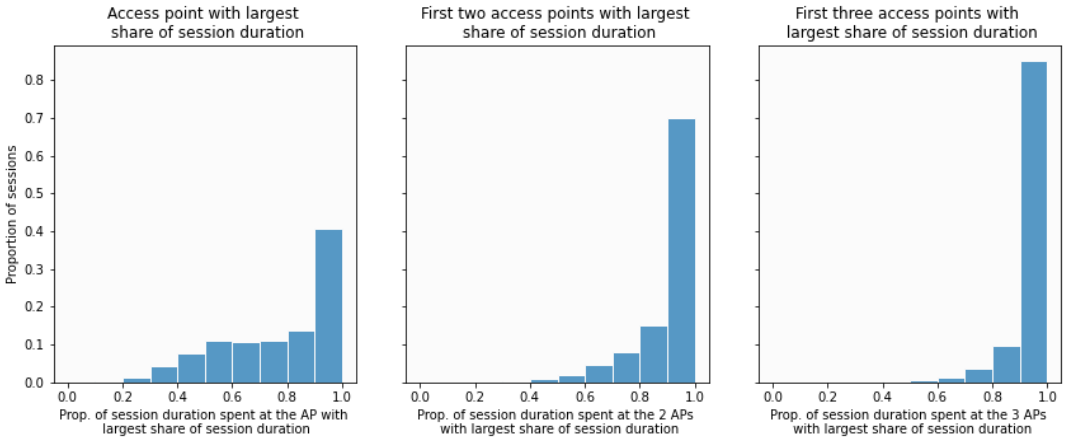

Figure 6.1: The proportion of the session duration spent at the access point with the largest share of the session duration, at the top two and the top three access points with the largest share of the session duration, for each session

points is limited to three, to avoid having a large number of access points as dominant in cases where no access points stand out. For example, if a user is not settled at any point and has connected to 10 APs, each with roughly 10% share of the session duration, it is better if only three of these access points are considered dominant, to prevent the model from classifying the entire session as 'Settled' (More details of how the model uses dominant APs in Section 6.3). This results only in a few sessions where the dominant access points to be associated with less than 80% of the session duration. However, the number of sessions where this happens is very small, as shown in Figure 6.1.

6.2 Adaptive Window

During a session, in order to predict whether the user is settled or not at a certain time, different activity metrics are calculated. These metrics are calculated for some time window preceding the timestamp for which the prediction is made.

Akbar et al. [1] proposed a prediction algorithm for IoT data streams called *Adaptive Moving Window Regression*. In their proposal, as data arrives in real time, the size of the moving window of data used for training changes depending on the value of an error metric. If they encounter high error, the window size decreases, otherwise it increases. In their case, this ensures that high prediction error is not propagated.

We follow a similar approach, but instead of using an error metric, we use the volume of events to determine the size of the window. If the volume of events is low, a greater window ensures that any noise during a long period of low activity does not have a great impact on the classification result. On the other hand, in periods of high activity, a smaller window ensures that the model is responsive and can quickly detect the start or the end of a settled period.

The size of the window is calculated using the number of events in the five minutes

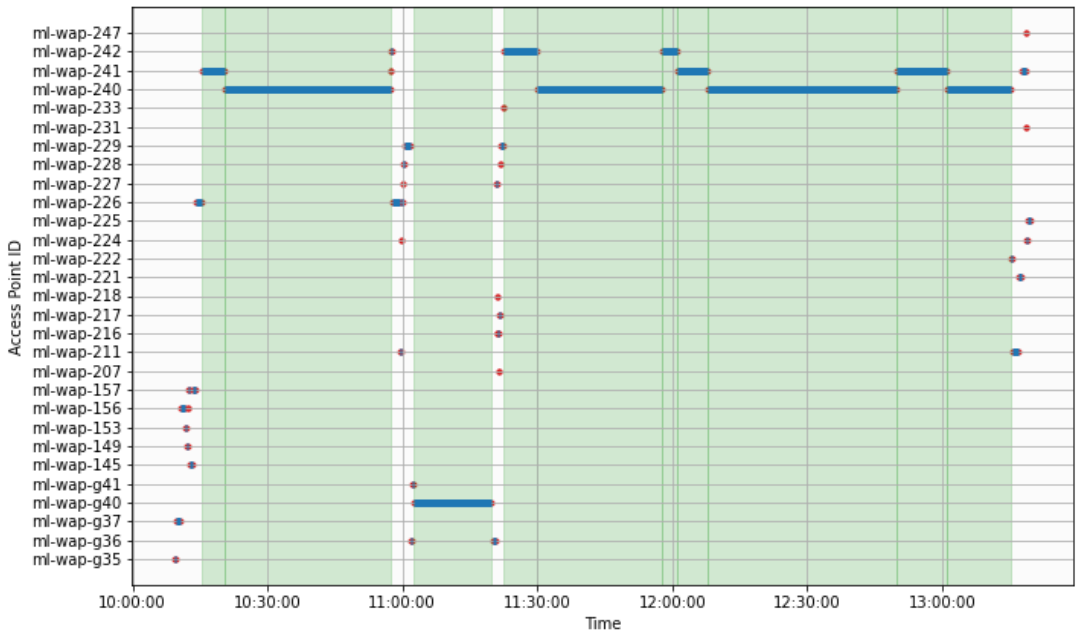

Figure 6.2: An example of a session with the time periods when the user appears to be settled at some area in the Library highlighted in green

preceding the timestamp when the prediction is made. The window size is initialised as the maximum window size. If a new event is recorded, the window size is halved. To ensure that changes to the window size are temporary, the window size is doubled if more than five minutes pass since an event was recorded. Therefore, the window size w for a session at time t can be expressed by the following equation:

$$w(t) = \max\left\{\frac{w_{max}}{2^{C_{[t-5,t)}}}, w_{min}\right\}$$

where w_{min} and w_{max} are the lower and upper bounds for the window size respectively and $C_{[t-5,t)}$ is the number of events recorded in the session in the 5 minute period preceding t, excluding t.

Figure 6.3 compares the value for one metric which is used for classification using a constant window size of 20 minutes against using an adaptive window with size between 5 and 30 minutes. The metric is the number of unique access points visited during the window (excluding the current access point) divided by the window size and it is calculated for the session given in Figure 6.2. More details about the features that will be used for classification are given in Section 6.3. From this example we can see that the effect of the adaptive window is that it amplifies changes in the metric when a large volume of events is recorded, such as just after 11:00 in the example. On the other hand, it dampens the impact of individual events on the metric when a low volume of events is observed, such as the events recorded around 12:00 in the example.

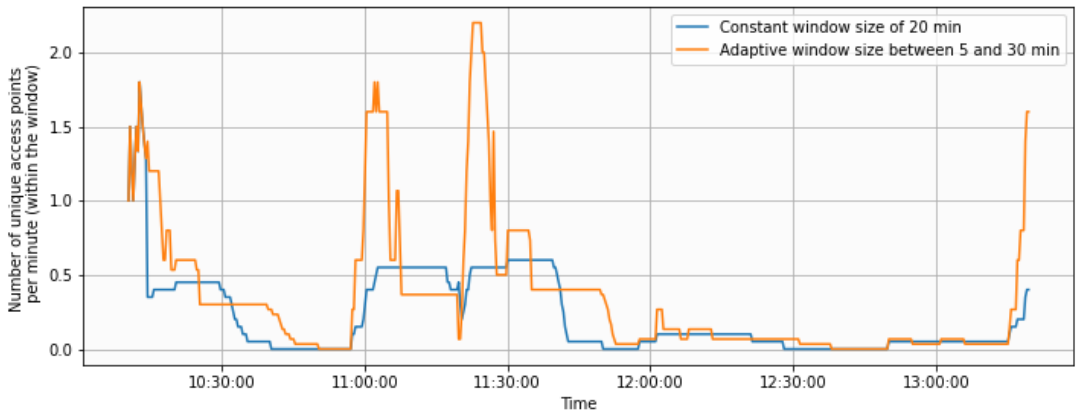

Figure 6.3: The number of unique access points visited during the window (excluding the current access point) divided by the window size over time for the example in Figure 6.2, calculating using both a constant window size of 20 minutes and an adaptive window with size between 5 and 30 minutes

6.3 Features for classification

Our goal is to classify events as '*Settled*' or '*Not settled*'. We refer to a sequence of events that are *Settled* as a '*settled period*'. To achieve this, we manually labelled the events for 33 sessions (531 events in total), to the best of our abilities in the absence of knowledge about whether the user was actually settled. We took into account the distance between access points (for APs with known locations) as well as the timestamps of each event to judge whether the user moved or not. For example, in Figure 6.2 the shaded regions show which parts of the session were marked as settled periods.

As a user moves within the library, their device connects to multiple access points, remaining connected to each one for a very short period of time, often less than a minute. This also means that they visit multiple access points in a short time span. Additionally, when events across multiple floors are recorded during a short period of time, it is unlikely that the user is stationary. We also assume that in the vast majority of sessions, especially in those where the user is seeking a study space, the proportion of the session where the user is settled far outweighs that where the user is moving. Therefore, if during the window preceding an event most or almost all of the window was spent at the dominant access points that are determined in the way described in Section 6.1, we can reasonably say that the user is settled. This is an important feature, especially in cases where the user can alternate very frequently between these dominant access points despite possibly not moving. This is something that has been observed in many examples of sessions where the user appeared to be settled in the Lower Ground floor of the building.

For each event, the corresponding window ends at the time when the next event is recorded and has size which is calculated in the way described in 6.2. Taking the observations above into account, we classify each event by calculating the following features using the window:

AP Count per minute: Number of unique access points visited during the window divided by the window size. The current access point is excluded to ensure that when no transitions to other access points are observed, the value is zero. Figure 6.4 shows the distribution of this feature, distinguishing between the two classes for the events in the 33 labelled sessions. Higher AP counts mean less likelihood of the user being settled in the time interval between the current event and the next event (i.e. during the current event)

Floor Count per minute: Number of unique floors visited during the window divided by the window size. The current floor is excluded to ensure that the value is zero when there are no transitions to other floors. Figure 6.5 shows the distribution of this feature. Higher values increase the likelihood of the user not being settled. The likelihood of the user being settled drops more sharply as the value increases compared to the AP count feature.

Proportion of window at dominant APs: The proportion of the window where the device was connected to any of the dominant access points. Figure 6.6 shows the distribution of this feature in the manually labelled dataset. The higher the proportion, the more likely the user was settled during the event. It is also interesting that for the vast majority of events, either for the entire window the device was connected to dominant APs, or to non-dominant APs.

Proportion of brief events per minute: The number of brief events recorded divided by the window size. We consider an event as brief if its duration (i.e. the difference between the timestamp of the next event and its timestamp) is at most 30 seconds. Figure 6.7 shows the distribution of this feature with lower values resulting in a higher likelihood of the user being settled during the event.

As noted above, for all four features there exist values which can separate the two classes, which we expect to be beneficial in obtaining good results from the classifier.

One other feature that was considered was the proportion of time spent at study spaces. However, this was not used as not all access points which are at study spaces are known (due to the incomplete floorplan information - see Section 4.1.3). This could result in the model missing settled periods at study spaces that are unknown. We also considered using the floor of the AP as a feature, as in different floors we observed differences in how many events are recorded and how often. For example, in the basement devices appeared to transition between more access points more frequently while appearing to be settled. These differences could be attributed to different equipment used or different layouts of access points in different floors. However, this would require a much larger amount of training data given that the library has eight floors. Therefore, this feature was not used at this stage.

6.4 Model selection

The classifier is trained using the events from the 33 manually labelled sessions - 537 events in total. Using 5-fold Stratified Cross-validation, the performance in terms of Accuracy and F1 score was calculated for different machine learning algorithms.

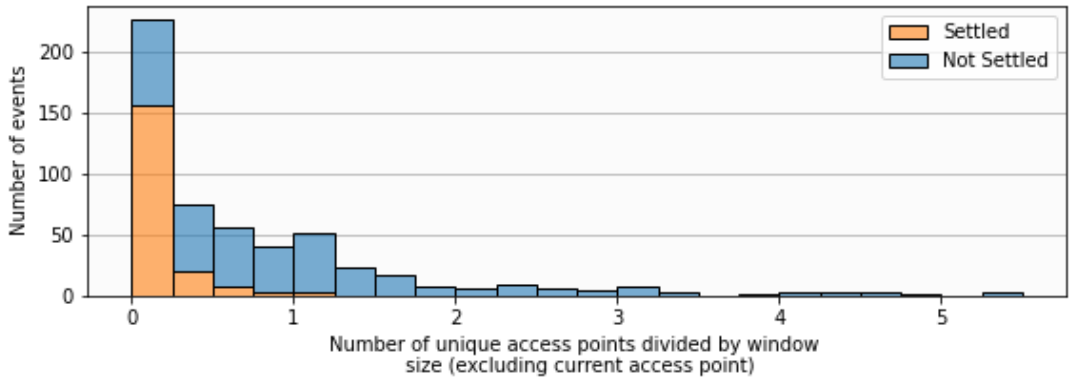

Figure 6.4: The distribution of the 'AP Count per minute' feature, distinguishing between the two classes

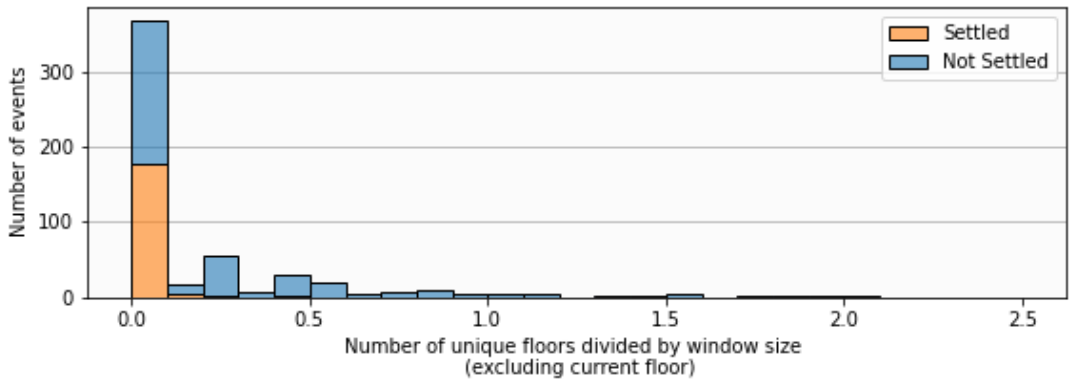

Figure 6.5: The distribution of the 'Floor Count per minute' feature, distinguishing between the two classes

Cross-validation is used, instead of simply splitting the labelled data in one training and one validation set, in order to avoid overfitting the model on that specific validation set when the hyperparameters of the model and the window size range are adjusted. The algorithms that we evaluated are:

- Logistic Regression
- Support Vector Machine (SVM)
- Decision Tree Classifier
- Random Forest Classifier (RFC)

The Scikit-learn [10] implementation of these models was used. Some of the hyperparameters of these models were adjusted in an effort to improve classification performance. For the SVM, the regularisation parameter was varied, for the Decision Tree the maximum depth of the tree and for the Random Forest the number of estimators, as well as the maximum depth. Table 6.1 summarises the results from these experiments, showing the accuracy and the F1 score for each model. Only the results using the hyperparameters and the adaptive window size range that yielded the best F1 score

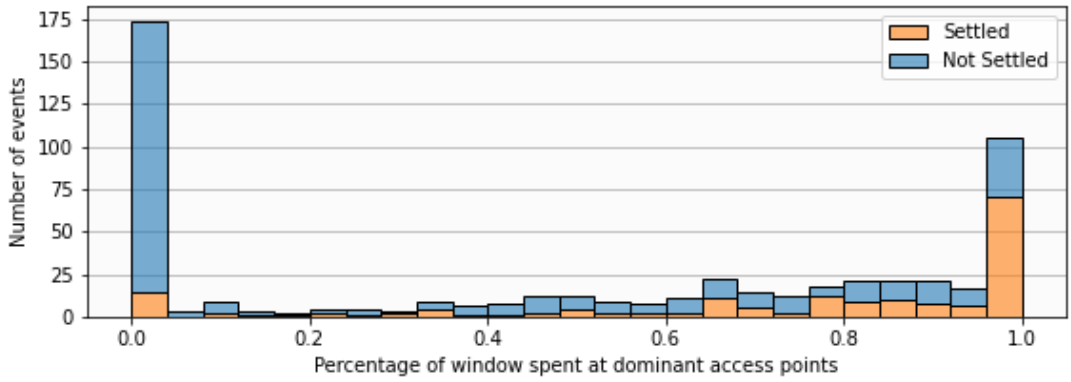

Figure 6.6: The distribution of the 'Proportion of window at dominant APs' feature, distinguishing between the two classes

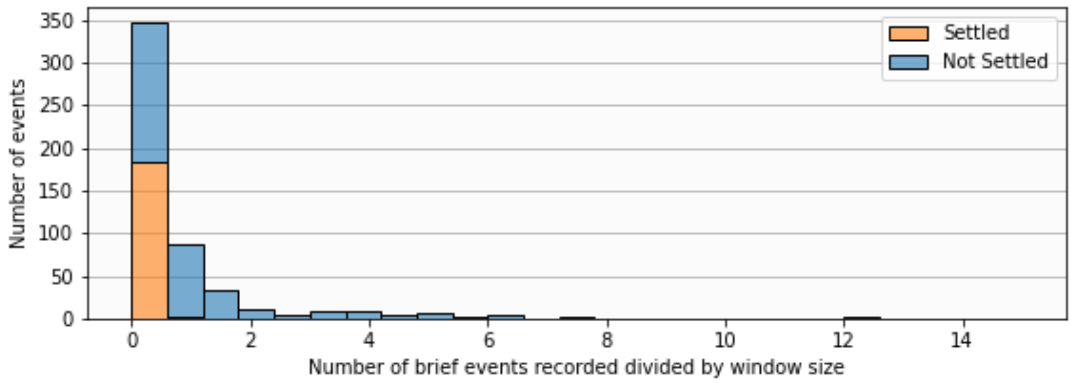

Figure 6.7: The distribution of the 'Brief events' feature

for each algorithm are shown.

The Random Forest Classifier with Maximum depth 6 and 100 estimators gave the highest scores both in accuracy and F1. Therefore, this is the model that was selected and which will be evaluated in Chapter 7. While the Random Forest provides the best performance and is selected as the winning model, the other models are not too far behind. Therefore, they could be taken into consideration in future work if a more robust way of labelling the data can be used.

The number of estimators is the number of trees that the Random Forest classifier grows. Figure 6.8 shows the F1 score on predictions made on both the validation and training sets for different values for the number of estimators. The maximum depth of the trees is kept constant at 6. As before, the F1 scores are calculated for each of the 5 experiments in the 5-fold cross-validation procedure and then averaged to give a single score for the training set and a single score for the validation set for each value of the number of estimators. A small value (e.g. 5) for the number of estimators results in lower performance in both the training and validation sets, suggesting that the model is underfitted. As the number of estimators increases, the F1 score increases for the training set while the score for the validation set peaks at 100 estimators. This could

Learning Method	Hyperparameters	Window Size	Accuracy	F1 Score
Logistic Regression	Regularisation parameter 1	[5, 7.5]	0.83	0.78
Support Vector Machine	Regularisation parameter 10, Linear Kernel	[5, 7.5]	0.86	0.80
Decision Tree	Max depth 6	[5, 30]	0.87	0.81
Random Forest	100 estimators, Max depth 6	[5, 30]	0.88	0.82

Table 6.1: Performance for each learning method, using the window size and hyperparameters yielding the highest F1 score

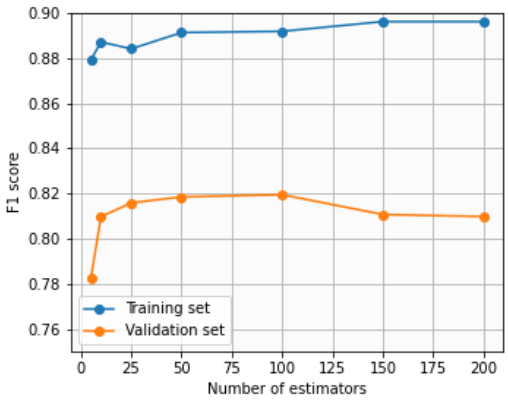

Figure 6.8: The effect of varying the value of the number of estimators of the Random Forest Classifier on the F1 score on predictions made on both the training and validation sets

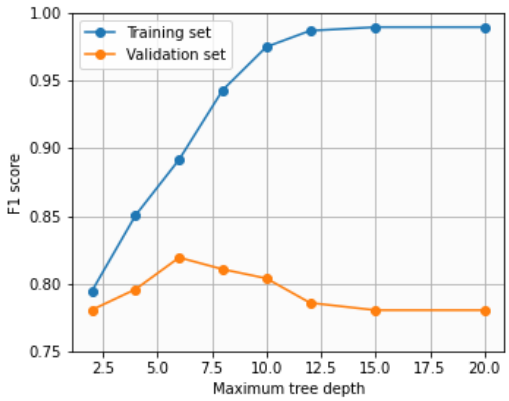

Figure 6.9: The effect of varying the value of the maximum tree depth of the Random Forest Classifier on the F1 score on predictions made on both the training and validation sets

suggest a little overfitting for higher values of the number of estimators (over 150). As the F1 score is maximum at 100 estimators, we pick that value for the hyperparameter.

Figure 6.9 shows the F1 score for different values of the maximum tree depth for both the training and validation sets. The maximum depth is the maximum depth each individual tree in the random forest can have. As shown in the figure, it has a more significant impact on the classification performance compared to the number of estimators. Very small values for the maximum depth (less than 5) result in low F1 score for both the validation and training sets, which could suggest that the model is underfitted. On the other hand, very large values (more than 12) of the maximum depth can lead to overfitting as the performance on the validation set drops, while the F1 score for the training set approaches 1 as the maximum depth increases. The F1 score for the validation set peaks at maximum depth of 6 which is the value we selected for our model.

Chapter 7

Evaluation

The model has been selected (RFC) and its hyperparameters have been adjusted using data from Dataset B (Max depth 6, 100 estimators). Now the model is trained on the entire training data (without splitting to training and validation sets). Dataset A has been set aside to be used for evaluation. We follow two different approaches for evaluation. In the first, in Section 7.1, the model is evaluated using a small sample of manually labelled data from Dataset A. In the second, in Section 7.2, a classifier is built that uses the output of the settled periods model to predict which access points are located at study spaces. We expect that users are settled for longer at study spaces and that more people settle at study spaces. Therefore, if the output of the settled periods model is sensible, a simple classifier should be able to use it to predict which APs are at study spaces with high accuracy.

7.1 Evaluation using a test set

Twelve sessions in Dataset A have been manually labelled and preprocessed in the same way as the training data, and the features are calculated as described in Chapter 6.

For the events in the testing set, the model gives 82% accuracy and 77% F1 score, which is slightly lower than the performance seen on the validation set in Chapter 6. However, the difference is not too large, especially considering that the size of the testing set is small. Therefore, it appears that the model generalises well.

Figures 7.1 and 7.2 show examples, one session in each, of the actual settled periods, as they have been manually labelled and the ones that are predicted by the model. In the first session, the predicted settled period matches exactly the actual one. It appears that the model handles instances of where the device alternates between multiple access points well, possibly thanks to the use of the dominant access points as described in Section 6.1. However, in the second session (Figure 7.2), the last settled period (starting just before 12:40) appears to have been missed by the model. It is possible that the model is less likely to detect a settled period if it does not take place at the dominant access points. Potential improvements on this approach could be explored in future work as described in Section 8.2.

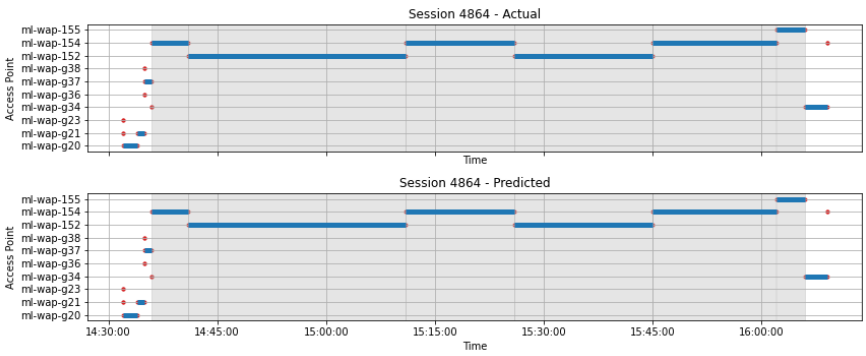

Figure 7.1: An example of the actual and the predicted settled periods for a session. The settled periods are shaded.

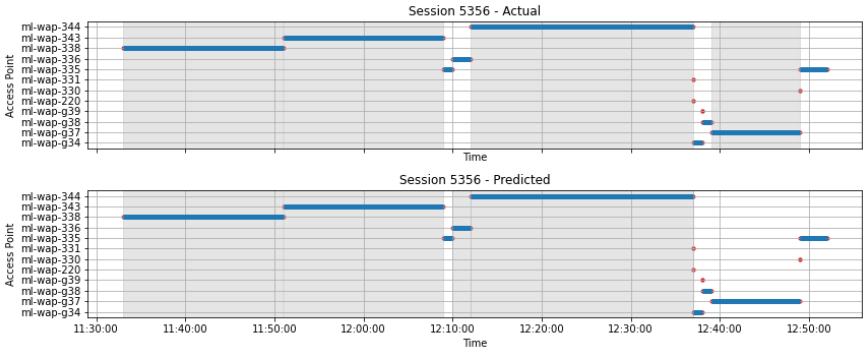

Figure 7.2: An example of the actual and the predicted settled periods for a session. The settled periods are shaded.

7.2 Evaluation by classifying access points

In order to evaluate our model we use its output to build another model that can predict a feature that is already known and calculate its performance. For access points for which their location is known, one such feature is whether they are located at a study space or not. Using the floorplans we can label some of the access points, indicating whether they were located at a study space or not.

We use 2500 sessions from Dataset A which has been set aside for evaluation. These sessions are preprocessed, the features are calculated using a window of size between 5 and 30. Using the successful model from Chapter 6, i.e. Random Forest with maximum depth 6 and 100 estimators, each event as classified as *Settled* or *Not settled*. Adjacent *Settled* events are combined to continuous settled periods. Each settled period is then matched to the access points the device was connected to for at least 30% of the settled period's duration. This means that multiple access points could be assigned to the same settled period.

For each access point (AP) we calculate the following features:

- Number of settled periods assigned to the AP.
- Number of settled periods assigned to the AP that are at least 30 min long.

- Mean settled period duration.
- Standard deviation of the settled period durations.

Our expectation is that, typically, users are settled for longer at study spaces, and the time for which they are settled varies more compared to areas which are not study spaces. This is verified by the results of our model, as shown in Figure 7.4 where most access points at study spaces have higher settled period duration mean and higher standard deviation. We also expect that more people settle at areas which are study spaces, and most of them are settled for periods of time at least half an hour. This is consistent with the results of our model. Figure 7.3 shows that the number of settled sessions at study spaces is higher compared to access points that are not at study spaces. In addition, the points corresponding to study spaces lie closer to the identity line, suggesting that a high proportion of their settled periods are at least 30 minutes long.

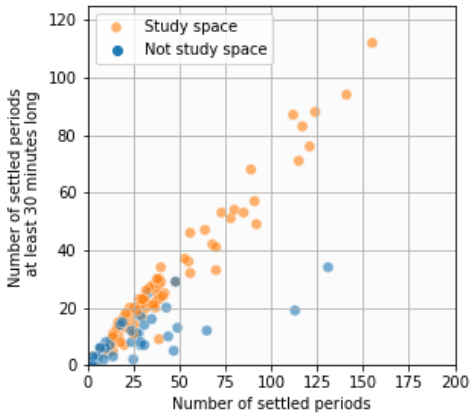

Figure 7.3: Scatter plot showing the number of settled periods and the number of settled periods which are longer than 30 minutes for each access point

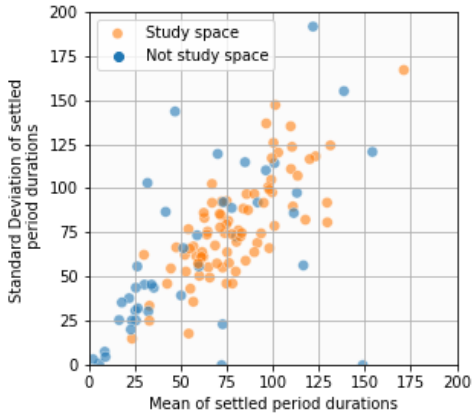

Figure 7.4: Scatter plot showing the mean and standard deviation of the settled period durations for each access point

We use Stratified 5-fold cross-validation to split the access point data in 5 folds. In each of the five experiments, one fold is used for testing and four for fitting the model. In each experiment, we fit a Logistic Regression model and use it to predict whether each of the access points in the testing set is at a study space or not. The performance metrics are calculated and averaged across the five experiments.

The result is 81% accuracy, 84% recall and 88% precision. This is a good result, considering that sometimes access points cannot be separated perfectly into study spaces or non-study spaces, as users can study in areas of the library that are not study spaces or some access points are near areas that are study spaces, as well as areas that are not. All study APs that were misclassified have a low number of settled periods, which is something common for non-study access points. However, for less busy study spaces this could mean that the values for the mean and standard deviation of the settled period might not be reliable.

The good results in predicting whether access points are study spaces suggest that the output of the model predicting the settled periods is sensible.

Chapter 8

Conclusion

To conclude this report, the main achievements of the project are summarised (Section 8.1) and some short-term and long-term future work towards the ultimate goal of monitoring the usage of study spaces and providing useful information to students and staff is discussed (Section 8.2). Finally, we comment on some key challenges that were faced and on lessons learned (Section 8.3).

8.1 Summary of work

The main achievements of this project can be summarised as follows:

- Establishing a procedure for obtaining access to data from the Main Library. This is the first time the Library has collaborated with someone from outside the Library. Therefore, a process which ensures the privacy of the users was agreed, which sets a good precedent for future projects and collaborations between the library and students.
- As part of initial plans for recruiting human participants, preparation has been made for future evaluation involving human participants, including a Participant Information Sheet.
- The available data from the University's Main Library (Wireless Access Point log, Turnstile log and Floorplans) were explored and their main features were summarised
- The possibility of whether an individual is identifiable from the data that was available to us was explored and it was concluded that it is very unlikely
- A pre-processing pipeline for the available data was constructed with the necessary justification and its limitations outlined. It involves removing noise, splitting the Wireless Access Point log into sessions, discarding outlier sessions and determining which device for each user is primary so the corresponding WAP log can be used as input to the model.
- A model that detects when a user is settled somewhere in the library (e.g. a study

space) has been constructed using a Random Forest Classifier. Four candidate models were compared and their performance was calculated using Stratified k-Fold cross-validation. The most successful one was selected. Cross-validation was also used to tune the hyperparameters of the model.

- The developed model was evaluated using a manually labelled testing set, giving accuracy of 82%, in terms of the proportion of events classified correctly.
- Another evaluation method that was pursued was building a Logistic Regression model that uses as input information from the settled periods detected by the previous model to predict whether an access point is at a study space or not. As we expect that students tend to be settled for longer periods of time at access points and more students settle at study spaces, the good performance of this model is a further indication that the output of the settled periods model is sensible.

8.2 Future Work

As outlined in Section 8.1, significant progress has been made towards the goal of monitoring the usage of student study spaces. The project can be expanded further both through better evaluation with physical access to the library and by additional functionality.

One of the major limitations in this project was inability to have physical access to the library due to pandemic-related restrictions, preventing access for purposes considered non-essential. This meant that our plans for recruiting human participants could not go ahead. As a result, this should be carried out in the future at the earliest opportunity. It will enable us to match patterns observed through the data to real actions of users. This will be useful for evaluating the correctness of our preprocessing steps and the performance of our model, ensuring that our assumptions are correct and enabling us to extract further insights from the anonymised data. In addition to the recruitment of human participants, we can conduct experiments in the library ourselves (e.g. by moving in a specific way) and check whether the resulting WAP log matches what we expect.

Even though the model appears to work well overall, its performance could be improved by developing an algorithm that can cluster several access points together. If these clusters of access points are treated as a single access point, noise could be reduced even further, improving model performance.

The data on the location of access points (floorplans) was outdated for some access points, and some were absent altogether. With physical access to the library we will be able to detect and map the precise locations of all access points. Then, we can experiment with using the physical locations of the access points for preprocessing or for our model. For example, the distance travelled in a given time could be estimated, which could be used as a feature in the model.

With access to data across several days, we could experiment with forecasting weekly or daily trends for different metrics, such as the time needed to find study space or the occupancy levels for specific parts of the building. The time needed to find a study

space could be extracted by combining the results from the model detecting when a user is settled at some area in the library with the turnstile data.

To deploy the model, some work will need to be carried out to establish how often the model will be trained and how. In addition, the preprocessing pipeline will need to be adapted so events are preprocessed as they arrive. This has already been considered to some extent, as for example, the window in the predictive model is always before the timestamp for which the prediction is made. In addition, ways to update some parameters such as the dominant access points and the primary devices as new data arrives should be developed.

In the long-term it could be possible to explore how the results from the work that has been carried out can be communicated to Library users and Library management in a way that is clear, easily understandable and explainable.

Finally, even though we used data from the University of Edinburgh's Main Library, as the data itself was relatively simple, compared to data that is used in the literature (for example, involving signal strength), we could explore the possibility of adapting the results to different settings such as museums, conference centres or other university buildings.

8.3 General Remarks

This project has shown that even with very simple data (recording only time and location), meaningful insights can be extracted. Working with simple data, proved to be a big challenge as all relevant work that was studied used more fine-grained data, such as signal strength etc. However, this has also been one of the most important contributions of this project. Due to the simplicity of the data, the same approach could be applied more generally, not just in the Library, but elsewhere in the University, in museums or potentially for crowd management in large events.

Another key challenge was the inability to access the Library physically for mapping the locations of all access points and for evaluation with human participants, something that was initially estimated to be a big part of the project. Therefore, we had to pursue alternative evaluation methods and build our model and the preprocessing pipeline using only the completely anonymised data. This was a weakness of our approach, as we relied on manually labelled data under some assumptions for training and evaluating the model. However, once more reliable data can be collected from participants, the model can be easily adjusted to take these into consideration.

Appendix A

Data Request

This appendix includes the data request that was submitted to the data providers for approval which outlined was data was needed, how it was used and how we ensured the privacy of the users. The attributes included in the data we were given access to differ slightly from the ones specified in this request. Section 4.1 describes the attributes of the data we have access to.

Requested Data

(1) Raw Wi-Fi Logs

- Anonymised user identifier
- Anonymised device identifier (To distinguish between multiple devices used by the same user)
- Time of day & day of the week & some week identifier (Week identifier only needed to distinguish between different weeks)
- Whether it is a new session, a termination of the session or a continuation of a session
- Wi-Fi access point identifier
- User category (Staff/Student/Other)

(2) Turnstile data for the main entrances and exits of the main library

- Anonymised user identifier linked to the corresponding Wi-Fi log user identifier
- Time of day & day of the week & some week identifier (Week identifier only needed to distinguish between different weeks)
- Location of turnstile
- Whether it is entrance or exit
- User category (Staff/Student/Other)

(3) Information and/or diagrams on the locations of Wi-Fi access points and layout of the main library

- Metadata enabling mapping from a Wi-Fi Access Point identifier to its location within the Main Library
- Floorplans/diagrams for the layout of the building and locations of Wi-Fi Access Points

We are requesting a week's worth of data, which is sufficient for an initial exploratory stage, though more data are likely to be required in later stages of the project. The week the data was collected should be during the previous academic years, i.e. before pandemic-related restrictions were put in place. The data should be for both staff and students, therefore an identifier is needed to indicate which category each user is.

Purpose of using the data

Our aim is to explore the potential use of this data to understand how people move around the library and use its resources. The data will be analysed in an effort to detect usage patterns and extract key categories of users. We will then attempt to train a predictive model that anticipates movements and resource usage from visitors of the library. In particular, we will focus on the usage of study spaces by students, with the aim of providing indications or predictions of how long it may take to find an empty study space at any given time.

The end result can become a useful tool for students, giving live feedback to those seeking an empty study space (in addition to the existing simple traffic light system). It can also provide anonymised analytics and insights to the library management in terms of how people use it and inform resource planning.

As the project progresses it could be possible that comparisons between previous years (pre-pandemic) and now are drawn to see how the measures and the booking system have impacted movement patterns in the main library. It may also help validate compliance with the government's guidelines for social distancing.

Privacy

We aim to make use of the data in an ethical and privacy-oriented way. Usage categorisation and prediction will be attempted independently of the users' identities and the exact time of year. Hence, we are only requesting de-identified data with the timestamps altered so only the time of day and day of the week are known.

Ethics Review

We will take additional steps throughout the project to ensure the ethical use of the data through external review as follows:

- The project will undergo the Informatics ethics approval procedure.
- A preliminary analysis of the data will be undertaken during the initial stages of the project to determine whether any individuals (e.g. library staff using their offices) can be identified and the relevant sensitive data will be removed.
- A date will be set to review progress and findings with the Ethics group for the Main Library data approximately halfway through the project.

Bibliography

[1] A. Akbar, A. Khan, F. Carrez, and K. Moessner. Predictive analytics for complex iot data streams. *IEEE Internet of Things Journal*, 4(5):1571–1582, 2017.

[2] Cristian Chilipirea, Andreea-Cristina Petre, Ciprian Dobre, and Maarten van Steen. Presumably simple: Monitoring crowds using wifi. Jun 2016.

[3] Cristiano di Flora and Marion Hermersdorf. A practical implementation of indoor location-based services using simple WiFi positioning. *Journal of Location Based Services*, 2(2):87–111, Jun 2008.

[4] Shona Elliott. Angry students claim Edinburgh University is turfing them out of library while studying due to lack of space. *Edinburgh Evening News*, Jan 2020. `https://www.edinburghnews.scotsman.com/education/angry-students-claim-edinburgh-university-turfing-them-out-\library-while-studying-due-lack-space-1370087`.

[5] Aurelien Geron. *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, 2017.

[6] J. Krumm and E. Horvitz. Locadio: inferring motion and location from wi-fi signal strengths. In *The First Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services, 2004. MOBIQUITOUS 2004.*, pages 4–13, 2004.

[7] I. P. Mohottige, T. Sutjarittham, N. Raju, H. H. Gharakheili, and V. Sivaraman. Role of campus wifi infrastructure for occupancy monitoring in a large university. In *2018 IEEE International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 1–5, 2018.

[8] Ariful Mondal. 7-steps predictive modeling process, 2016. `https://rstudio-pubs-static.s3.amazonaws.com/223423_8ca6fccca1e44939be3f85ecbfa9598f.html`.

[9] Mohamed Ouf, Mohamed Issa, Afaf Azzouz, and Abdul-Manan Sadick. Effectiveness of using wifi technologies to detect and predict building occupancy. *Sustainable Buildings*, 2017.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos,

D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[11] Peter Savnik. Wi-fi positioning for crowd management, Jul 2016.

[12] The University of Edinburgh. Student factsheet 2014/2015. 2015. https://www.wiki.ed.ac.uk/download/attachments/102896991/Student_Factsheet_31072015.pdf.

[13] The University of Edinburgh. Student factsheet 2019/2020. 2020. http://www.docs.sasg.ed.ac.uk/gasp/factsheet/Student_Factsheet_31072020.pdf.

[14] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282 vol.1, 1995.

[15] University of Edinburgh Information Services. Main library occupancy traffic light system. 2018. https://www.ed.ac.uk/information-services/library-museum-gallery/using-library/lib-locate/main-lib/main-library-occupancy-traffic-light-system.

[16] University of Edinburgh Information Services. Main library floor occupancy indicator. 2020. https://www.ed.ac.uk/information-services/students/study-space/main-library-floor-occupancy-indicator.

[17] University of Edinburgh Information Services. Space occupancy monitoring. 2020. https://www.ed.ac.uk/information-services/students/study-space/space-occupancy-monitoring-pilot.