Final Year Project

Wireless Network Connections as a Proxy for Human Presence

Kevin O'Sullivan

Student ID: 16501573

A thesis submitted in part fulfilment of the degree of

BSc. (Hons.) in Computer Science

Supervisor: Professor Gavin McArdle

UCD School of Computer Science

University College Dublin

January 22, 2021

Table of Contents

Abstract

In the era of data driven business decisions and of ubiquitous mobile computing, human mobility prediction and location busyness prediction are fields that offer valuable insight to a diverse group of businesses, industries and service providers. This study uses a dataset of user associations with a WiFi network to create a model for predicting human presence. This study will firstly assess the current state of the art of the entire field through research and literature review. Based on this research an approach is defined, which includes the implementation of a number of machine learning algorithms ; Support Vector Machines, Decision Trees and Naive Bayes, followed by the evaluation of these models and selection of the most efficient model based on a a number of accuracy metrics ; Classification Accuracy, Precision, Recall, F1 Score and Mean Absolute Error. Using this model, a web application will be built, which will visualise the the model created, and will be overlayed on a map. The visualisation will have a drill down capability displaying a 24 hour summary of busyness of a location. The application's model and front end will be evaluated in terms of performance, usability and accessibility through means of performance metric analysis, user completed surveys and state of the art benchmarks.

Chapter 1: Project Specification

Given a dataset containing records of user association with nodes on a WLAN network, create an algorithm to predict the busy-ness of an Access Point (AP) on the WLAN network. In addition to this, create a web-based application that spatially represents the network and the network users over time.

Core :

- Through a thorough literature review, I will assimilate the current state-of-the-art in human mobility analysis and mobility prediction.
- I will spend time to understand the data.
- I will identify suitable features for predicting the busy-ness of a location, based on the data provided.
- I will develop an algorithm to predict the busy-ness of a location.
- I will evaluate the effectiveness of the approach.

Advanced :

I will develop an interactive web-based spatial dashboard to display the results.

Chapter 2: **Introduction**

In today's world, mobile device usage is ubiquitous and as a result of this, vast stores of mobile device generated data are available for analysis. These datasets contain everything we need to work towards accurately modeling human mobility.

The application potential for an accurate model of human mobility at any scale, whether it is a campus, large metropolitan area, national or international, is enormous. The demand for a model that reliably and accurately predicts human mobility can be seen across a broad spectrum of industries and fields, including but not limited to urban planning, epidemiology, network resource allocation [1], transport network design and disaster relief [2]. While some application areas for human mobility prediction are obvious such as smart bus route planning [3], there are also novel usages being explored for predicting human mobility as in [4] which uses interference in WiFi signals between nodes to predict how many people are in a room. There is much insight to be gained from accurately modelling human movement.

This study will focus on modeling human mobility in a defined local environment, specifically modelling human mobility on a university campus in KTH Royal Institute of Technology in Stockholm, Sweden. A dataset containing records of users associating with nodes (APs) on the university WLAN network will be used for this study. From this dataset features will be identified that can be used to create an algorithm that can be used for the prediction of the busyness of a location. As part of this process I will review the current state of the art in human mobility prediction and prediction of the busyness of a location. There have been many studies in this area and this study will take into account various approaches to similar problems. A model will be selected based on complexity, performance and accuracy. The end goal of this project is to create a web app capable of spatially visualising the model of location busyness across the network.

This field of human mobility prediction is an ever changing one, given the constant development of new technologies alongside the proliferation of mobile devices and networks there is always cause for exploration in the area and also for re-evaluation of current state of the art. The insights that can be gained by accurately modeling human mobility are powerful, and valuable to many companies and industries. This combination of capability due to new technologies , demand due to industry value, and the earlier mentioned readily available data sources is the reason this project is of value and very currently relevant.

The remainder of this report is organised as follows : Section 3 details related work in the area. Section 4 details the data considerations. Section 5 details the outline of the approach to the project. Section 6 is a work plan of how the project will be completed. Finally section 7 provides a summary and conclusions.

Chapter 3: Related Work and Ideas

There has been much work done in this area of human mobility prediction before. There have been many approaches taken to modelling human mobility, varying from Markov Predictors [5], to machine learning [6]. There have been studies regarding the privacy concerns of these approaches due to the invasive nature of modelling an individuals movement patterns. Predicting human mobility as a process is largely defined by the area you are predicting the movement within, there have been studies that have done this on international scales, and studies on a college campus [7]. The State of the Art will be assessed here in this section of the paper.

3.1 Mobility prediction in a local environment

Many studies have looked at modelling human mobility in a defined local environment. In today's world of smart cities, modelling the movement of a city's populus provides valuable insight for the allocation of resources across the urban area eg. emergency services while also being useful for intelligent bus route planning [3], traffic forecasting and many other applications. In many traditional mobility modelling studies including [8],[7] the prediction models are based on the context of the user's location history and current location. User trajectories are created at an individual level by abstracting from the data and a model of the user's mobility is constructed. This approach can be effective, however it does not take into account any temporal patterns ie. daily usage/movement patterns etc. A model using spatial-temporal combination may in some cases be more applicable, an example of which can be seen in [9]. This model considers the probability of a user moving to a certain location given the current time, as well as given the user's location context.

3.2 Different predictors

Various studies into human mobility prediction have tried many different methods. There are many different situations in which modelling human mobility is done, from university campus to taxi demand prediction to assisting with post disaster resource provisioning. Due to this variety in the landscape many different methods for predicting and modelling the humans in question have been used. Many studies take a machine learning approach, which is outlined in more detail in the next paragraph. Methods for estimating human occupancy or the number of people in an area date all the way back to simply counting people manually that are attending events or using sensors to detect the presence of people. In terms of more advanced model building there are many techniques and tools available. A number of other studies including [5] and [9] use markov predictors. Markov predictors make a prediction based on the current user context and user history. These studies use the users location history as context. In other studies like [7] the patterns of users on the network is analysed statistically and modeled as a time varying poisson arrival model. In [7] the arrival of users on the network can be seen to happen in two stages throughout the day and the distribution is modelled as [7] two stage hyper exponential distribution. Another approach that has been used is Neural Networks. In [10] a number of neural network based approaches are

considered, involving pre-trained neural networks and untrained ones.

3.3 Machine learning for mobility prediction

Recent advances in Machine learning have made available a wide variety of new and powerful tools for data analysis, modelling and prediction. The problem of predicting human mobility requires some user context. This context can be the user's location, current time etc. For this reason the problem of human mobility prediction is suitable for machine learning to be used as a solution. Machine learning treats the problem as a classification problem and classifies user movement based on the user context provided. In [11] an extensive study was carried out to assess the various machine learning approaches to modelling human mobility. It found that these approaches could be largely categorized into three categories; user modelling, place modelling, and trajectory modelling. It further divides these approaches into supervised and unsupervised machine learning approaches. [11] differentiates these approaches based on a number of factors: "type of tracked object, trace time resolution, and resolution and accuracy of the location positioning technology". In this study we seek to model a university campus and this falls under the category of "place modelling" and this study will also use a supervised learning approach. Another study [12] assesses the ability of a number of machine learning algorithms to predict human mobility and found that an ensemble-learning algorithm combining the votes of 4 base classification models (Bayesian Learning, Decision Tree Learning, Rule induction learning and instance based learning) was the most accurate. A very interesting study [6] compares a number of different machine learning approaches at their ability to predict the number of occupants in an office building based on a dataset of environmental sensors, and then also on a WiFi trace set. This is similar to what this study seeks to do, especially when considering the WiFi trace set. In particular [6] compared Support Vector machines (SVMs) , k nearest neighbours (kNN) and Artificial Neural Network (ANN). It found that over the whole period, using only the WiFi trace dataset that SVM resulted in the lowest MAE, MAPE and RMSE error scores. Another similar study [13] using a WiFi trace set from an office block compared ANN, Random Forest and LTSM (Long term short term memory networks) and using RMSE as error metric found that Random Forests were the most accurate for predicting occupancy of the building. [13] also had some interesting data considerations such as not using devices that were continuously connected for the entire duration of the study as they do not infer occupancy.

3.4 Limit of human predictability

When working to create a model of user movement in an area, it important to keep in mind that human mobility is a stochastic process. It has a certain level of randomness to it as humans make decisions based on many things. In [14] the entropy of users movement was calculated and the limit of predictability was found to be very high at 93%. However another study [15] discusses that by changing the degree of granularity of the spatial data or the temporal data the limit of predictability is reduced. It is affected by a number of variables.

3.5 Busyness of a location

This study is focused on predicting the busyness of a location specifically on a college campus and so it is not necessary to model human mobility at an individual user level, but instead to understand it at an aggregate one. The busyness of an AP can be viewed as a count of users connected to an AP on the WLAN network at any given time. Similar studies and projects to this have been done before. This count of people can go from a very basic manual count of people at an event, or using sensors in a system to count cars in a car park, or as we are doing in this study, using WLAN connections to count users connected to an AP. In [16] a parking guidance system uses a large system of sensors in parking lots, and a management system that gives feedback to the system users on parking availability. This is at its core a system to give feedback on the busyness of its location. Each parking spot is fitted with a sensor and these sensors cumulatively count the parked cars, which in turn provides a busyness classification. In [17] a study inferring land usage form mobile phone data from cell tower connections, a count of how many users are in each urban zone is required. To get this count an aggregation of all the connections per tower is used. This study will use a similar method to achieve an aggregate count of how many users are connected to each AP on the WLAN network. This avoids any privacy concerns and is also adequate for calculating the busyness of a node on the network, as individual movement patterns are not required for this. In [18] a study predicting the flow of users between grid squares, a dataset of mobile phone network data is aggregated at a grid level is used. The data is aggregated into grid squares, with each squares having a count of users in it at any point in time, it then uses these to estimate the flow of users between squares on the grid. A very interesting study [4] with a novel usage idea for WiFi attempts to use WiFi signals to estimate the number of people walking in an area. The people are not required to have any WiFi capable device. The system works based on interference in the signals going between the WiFi nodes caused by the people walking in the area. The system analyses the blocking of the 'LOS' (Line of Sight) between the two WiFi nodes. This is an interesting application of WiFi to estimate the number of people in a room or area and could prove powerful in the future. Its lack of requirement of WiFi enabled devices by the tracked person is also very interesting. In [19] a model of human occupancy of buildings on the MIT campus is created. This study uses WiFi trace data in the same way as this project intends to. A dataset of WiFi activity from about 5,000 WiFi hotspots was used. This model provided an overview of the number of people in different areas of the campus, and can be used at a building level also. It is noted in this study that with the full deployment and installation of a system of sensors around the campus a more accurate model could be created, however this model based on WiFi activity has no pre-requisites and is very low cost in terms of financial cost and time cost. Another study monitoring passenger usage of transport routes via WiFi connection can be seen in [20]. The study keeps track of users by the MAC address of their device and so can track recurring users and identify user travel patterns and duration. Another study [21] created real time visualisations of human occupancy in a library based on WiFi and Bluetooth device detection. It found that it was far better to use WiFi because many people do not have Bluetooth active on their device, with a much higher detection rate found through WiFi scans. It is common for large facilities to use CO2 detectors to approximate the building occupancy for the purpose of managing HVAC (Heat Ventilation and Air Conditioning) systems efficiently. In [22] the accuracy of using WiFi data to predict human occupancy is compared with CO2 detectors.An hourly scan of the WiFi network was done, and compared with the every ten minute CO2 detectors data, and it was found that the WiFi based count of people in the building was in fact more accurate. This is very interesting as it shows that WiFi being used as a proxy for human presence has genuine practical application and can be more efficient than current or old systems. These are some very interesting examples of other studies and experiments done using mostly WiFi detection to predict human presence and in turn, the occupancy or busyness of a building or location.

These studies come from the field of human mobility prediction and predicting human occupancy or presence. Some are very closely related to this study, and others more loosely related. Similarly to some of the studies this study will take a machine learning approach to the issue of predicting

the busyness of a location. This study also will be predicting the busyness of a location at a location level ie. by predicting how many users will be at a certain AP at a certain time, and so will not need to predict movements of users at an individual user level.

Chapter 4: Data Considerations

The dataset [23] contains records of authenticated user associations to the wireless network of the KTH Royal Institute of Technology in Stockholm. The dataset also includes scan results and mapping information of Wi-Fi networks, collected by means of war-walking at the university's two largest campuses. The dataset conatins 2.5GB of data. The data is structured as follows :

The records of user associations with APs in the network are stored in 16 .csv files, each representing a month of the 16 month period over which the data was collected. The time period over which the data was collected was from January 2014 to March 2015 inclusive.

Each record in these .csv files represents a user associating with a WiFi node on the WLAN network. These user association records have the following format :

timestamp, hashed_user_id, ap_id

Field Name	Description	Format
timestamp	Time at which the association happened.	dd/mm/yyyy mm:hh
hashed_user_id	Hashed Unique ID for user	Length = 40 chars
ap_id	Unique ID for Access Point	'Bldg' + alphanumeric sequence

There is also an APlocations.txt file that contains the locations of all APs.

Its columns are separated by commas and are as such :

AP, x_coordinate(m), y_coordinate(m), floor

Interestingly in [19] and [13], which are both studies using WiFi traces sets to estimate or infer human occupancy, they discuss data considerations relating to not all WiFi connections implicating human presence. For instance they do not use devices that are connected for 100% of the data collection time, as they consider these devices to likely be desktop computers or servers that do not necessarily infer human occupancy. [19] also considers that some users may have two devices connected. However, they do not take into account that some people that move through the area spanned by the WLAN network do not associate with the network, perhaps they use mobile data providers or do not use their device while in the area. For the purpose of this study it is considered that these effects will balance each other out.

Privacy Concerns

The data providers Ljubica Pajevic, Gunnar Karlsson and Viktoria Fodor created a coordinate grid on the university campus, and the coordinates provided in the data for the AP locations are based on this coordinate grid, rather than being latitude and longitude. This was done to avoid privacy concerns. They considered that some users of the network have very recurring movement patters .ie a professor of computer Science will always enter the computer science building before the lectures they give and leave after, and due to these potentially obvious recurring movement patterns. They also considered that building locations may be exposed through latitude and longitude coordinates. Without knowing the 0,0 coordinate (The origin point of the grid the coordinate system in question) it is impossible to overlay the APs on their exact location. Due to this, either an approximation

of the AP locations will be made, or the network will be overlayed on an arbitrary area, but the full functionality of the application will be demonstrated nonetheless.

As many approaches model human mobility at an individual level, there follows privacy concerns. In [24] it is estimated that in a dataset where user location is updated hourly and the spatial granularity of the set is equal to that provided by mobile network carrier antennas, just four spatial-temporal points are enough to uniquely identify 95% of individuals. In the modern world of technology there is almost no such thing as anonymity. Privacy concerns are higher than ever and many people consider these forms of surveillance to be far too invasive. For the purpose of this study an approach will be used to model the business of each location on the campus, and in turn the campus as a whole will be implemented, rather than creating a model that is built from individual user level. This approach minimises privacy invasion of users in the dataset. The user IDs in this dataset are also hashed which provides privacy.

Spatial Data

The 'APlocations.txt' file contains the locations of each AP in the WLAN network. These locations must be overlayed on a map for use in visualisation. Due to the fact that the coordinates provided are based on an unknown grid, the map used will be arbitrary.

The user association records must be aggregated by AP and a set time granularity of 1 hour. These aggregations will result in a count , per AP, of how many users connected to that AP in that hour. The machine learning algorithm will then be trained on these counts and will build a model.

Chapter 5: Outline of Approach

This study focuses on predicting the busy-ness of a location on a college campus. This will be predicted at a location level, and so the dataset of user associations with the network must be aggregated in such a way that busyness of APs at certain times are known. The first step is to get a count of how many users associated with each AP at hourly intervals for the entire period of data collection. The study will also experiment with different frequencies of aggregated counts eg. every 30 minutess / 90 minutes to assess the effect of the time granularity. These counts of the number of users at an AP will be classified ie Busy, Very Busy , Extremely busy, based on the min/avg/max counts of users at that AP historically. After this step the data is in the format required to train our model.

The training set for the machine learning algorithms will contain two predictor variables 'AP ID' and 'date time', and will contain one dependent variable 'location busyness'. Both predictor variables are categorical. Naive Bayes seems appropriate given it will classify the busyness based solely on the probabilities calculated from the busy-ness of the location historically. Decision trees are also an applicable algorithm here, however there are very few attributes to split the tree on. SVMs (Support Vector Machines) also are applicable here. Based on this the three most suitable machine learning algorithms are Naive Bayes, SVMs, and decision trees. This study will assess these three algorithms based on the accuracy of prediction and the most accurate algorithm will be used for the final model and application. The machine learning algorithms will then be tested using hold out testing, and assessed on its accuracy of prediction using a test set of data that was not used for training.

The final product of this project is a web application capable of spatially visualising the campus, displaying the busyness of each AP/Location. overlayed on a map. It will also provide a drill down capability on any AP/Location and this drill down will show an average distribution of the busyness of that location over a 24 hour period. The web framework used for the application was narrowed down to flask [1] or django [2]. These are the two most used python frameworks by far and are both suitable for this project. Flask is lightweight and good for small projects with a very defined scope, similar to this, for this reason it is flask will be used to develop the web app. The visualisation library used to develop the visualisation is narrowed down to one of two frameworks chart.js [3]or d3.js [4]. Both are very useful libraries, with d3.js having a more extensive set of capabilities. Chart js has a very straightforward API, but may not be as appropriate for this visualisation, especially concerning overlaying the chart on a map. D3.js will be used in the final implementation of the visualisation, as chart.js has a much narrower set of visualisation capabilities and it is not certain it would be able to implement the desired visualisation. The web app will use a bubble chart overlayed on a map of the area, where bubble size represents the busy-ness of the location.

The performance and usability of the app will be tested and assessed by a number of users by means of a Usability Questionnaire.

In Figure 1 below a wire frame mock up of the front end of the web app can be seen. The circular nodes on the map represent APs on the WLAN network. The size and color of each node both represent the busyness of the respective AP. The larger the node, and the more red the node is, the busier the location is. The key on the left of the visualisation describes the association

[1] http://flask.palletsprojects.com/en/1.1.x/
[2] https://www.djangoproject.com
[3] https://www.chartjs.org
[4] https://d3js.org

between the node colours, node sizes and the busyness levels. This system using both size and colour to represent the busyness enhances the immediate readability of the visualisation while also accommodating for colour blind users, therefore increasing accessibility as well as the usability of the application.

In Figure 2 below a wire frame mock up of the drill down capability of the visualisation can be seen. If the user selects a node, they will be presented with a simple 24 hour summary of the average busyness of that AP. Each bar represents an hour, the height of the bar and the color of the bar represents the busyness of the location.

Figure 1 : Wire Frame Mock Up of Web App Landing Page :

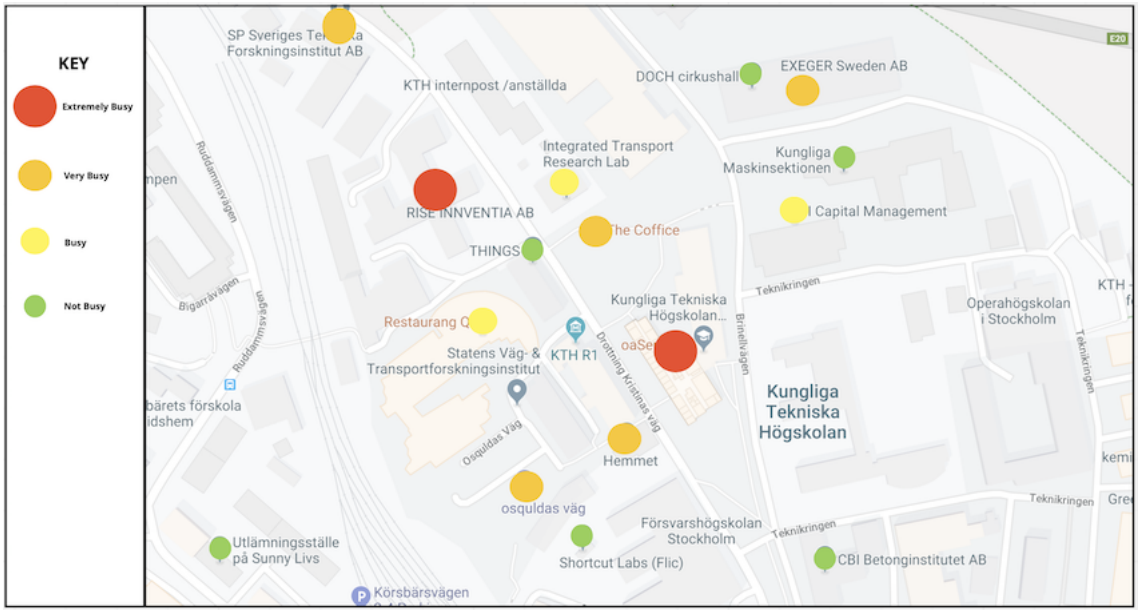
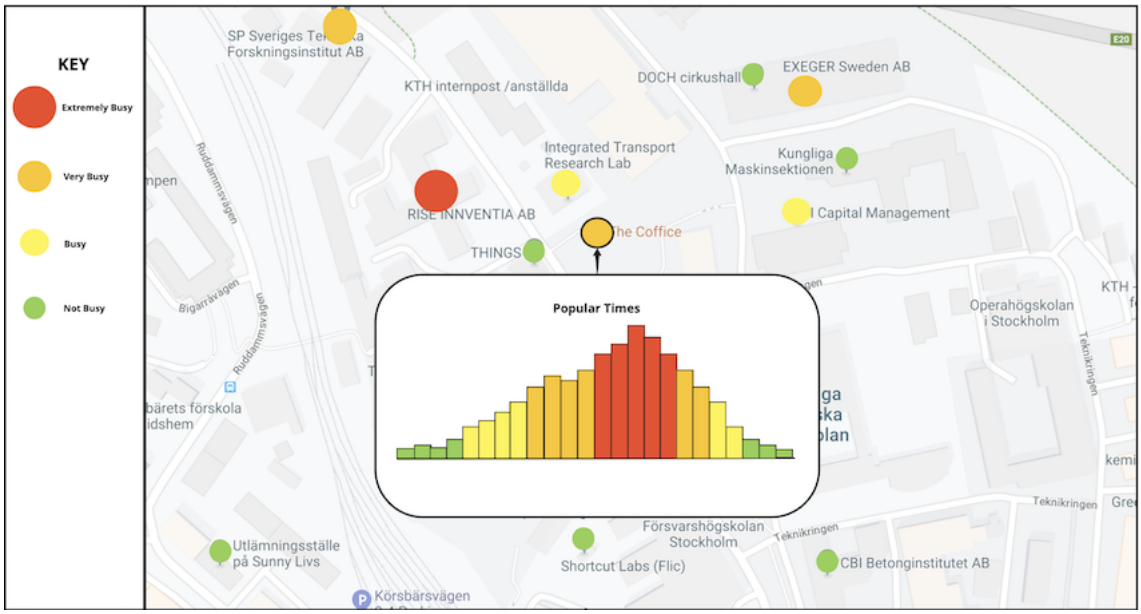
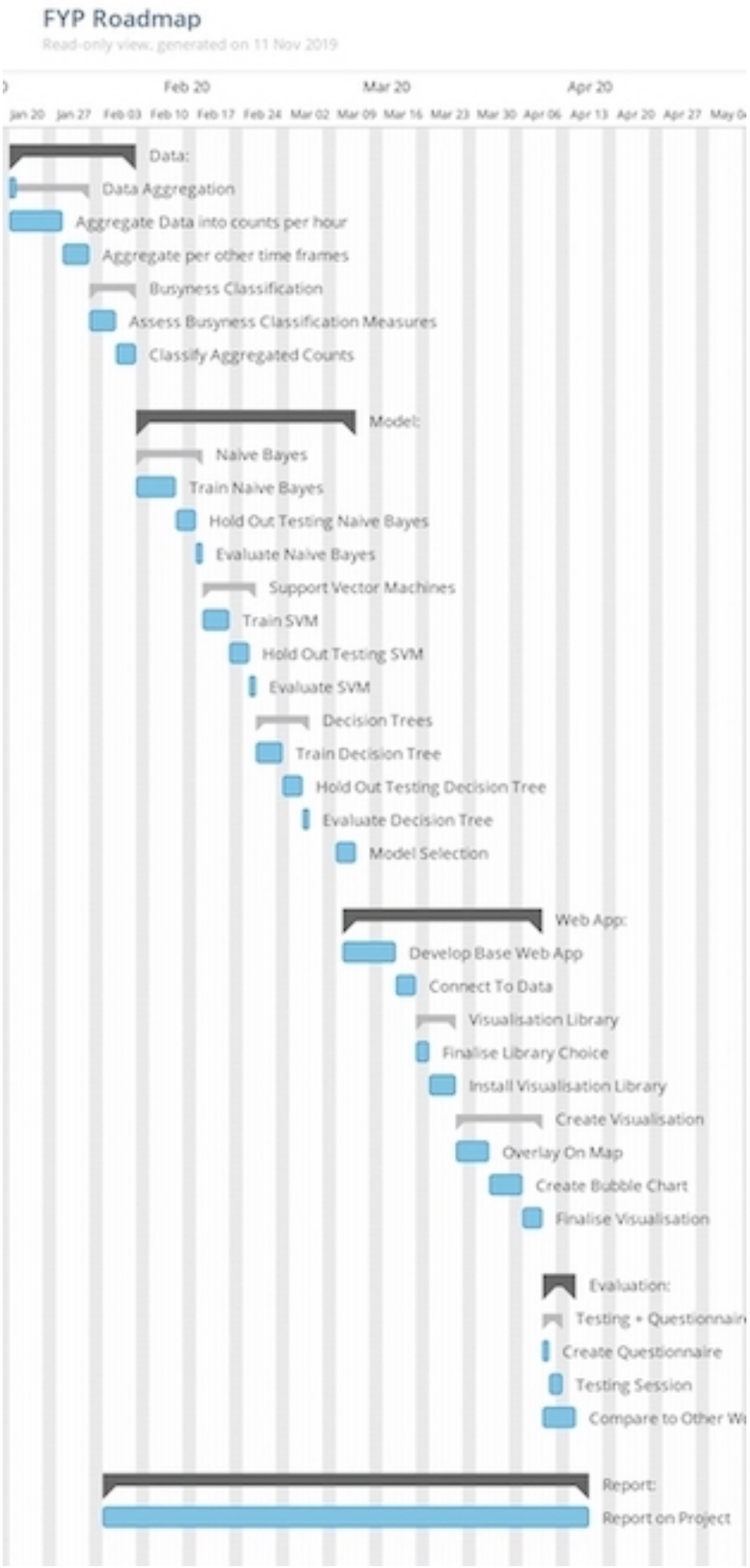


Figure 2 : Drilldown Capability Wire Frame :



Chapter 6: **Project Workplan**

Figure 1 : Project Gantt Chart



Firgure 1above shows a gantt chart of the of the development of the project. The tasks and

steps are detailed below. They are divided into four main sections : Data, Model,Webn App and Evaluation.

The planning of this project followed the steps of the CRISP-DM [1] methodology. CRISP-DM, which stands for Cross-Industry Standard Process for Data Mining, is a methodology for planning projects relating to Data Mining.

Data

The first stage of the implementation of this project is to prepare the data that will be used to train the machine learning algorithms. The dataset used contains records of users associating with a WLAN network, where as the machine learning algorithms will be trained on records of 'AP ID', 'Date Time', and 'Busyness' values. The first step here is to get aggregate the data by hourly count. To do this , at each hour of the entire period of data collection, a count of the number of users who have associated with that AP will be calculated. The distribution of busyness at each AP will then be assessed, and each AP will be allocated a scale on which it's busyness will be classified. Once this scale is created, the counts are converted to the busyness classifications (busy, very busy, extremely busy). Once this has been completed we have the data in the correct format for the machine learning algorithms to begin training on the data. The training set is in the form : 'AP ID', 'Date Time', 'Busyness'. The set of data must then be split into training and test data, in order to keep some data separate for test and avoid overfitting.

Model Training

In order to select a machine learning algorithm for this project, an assessment of 3 models must first be completed. These three models are Naive Bayes, Decision trees and Support Vector Machines. Each of these three models will be trained on the training data. Hold out testing will then be done on each model, and each model will be assessed using the following evaluation metrics : Classification Accuracy, Precision, Recall, F1 Score and Mean Absolute Error. Based on this initial evaluation, the model which performs best will be selected as the model used throughout the rest of the project. At this stage the final model has been identified, trained and tested using hold out testing.

Web App

The next stage of development will be the creation and development of a web app that will run the visualisation. A basic web app will be created as a basis for the rest of the development. With this web app set up, the next step is to connect the application to the static data source. Once this is complete the visualisation itself is the only outstanding piece of work in development stage.

The visualisation library must now be installed in the web application. D3.js will be installed in the application and then the creation of the visualisation begins. The first step of creating the visualisation will be to insert a map of the required geographical area (KTH University Campus) . This map will be the background of the chart. The next step is to create the bubble chart itself. This chart will be composed of a number of circles or 'bubbles' , whereby each bubble is overlayed on the geographical location of an AP within the campus, and the size of each bubble represents the level of busyness of that location. The final element of the visualisation to be completed will be the drilldown. If the user selects an AP, they will be presented with a 24 hour average of the busyness of that AP. Once this is done all components of development have been completed, and some final touches will be completed on the web app front end.

Evaluation

The final stage of this project will be the evaluation of the entire application. Firstly an evaluation

[1] http://crisp-dm.eu

of the machine learning model will be completed. This will include the analysis of the accuracy of the model in comparison to other similar projects, as well as an analysis of the accuracy scores discussed previously : Classification Accuracy, Precision, Recall, F1 Score and Mean Absolute Error. The final element of the project to be evaluated is the front end of the application in terms of its usability and performance. A testing session will be setup in which a number of users test the application, and afterwards fill out a questionnaire which will be collected and the results gathered and analysed. From the feedback received, a proposal of improvements and enhancements will made for potential future work. The effort and time required to implement these improvements and enhancements will be assessed and any feasible implementations will be made.

Risks Contingencies :

The work plan is detailed and has been thoroughly considered, and so should allow the required time for all development. However there are still a number of risks and contingencies to be identified with the project. The two main risks in the development process are the model training, and the web app development.Each model has been allocated approximately 1 week of time for training, hold out testing, and evaluation. There is a small risk that this time is not enough. The development of the web app is the other main risk in terms of time required. There may be some unseen obstacles, or some more time required to get the web app to the desired standard.These risks will not be a big issue however, as all sections of the project were allocated time which included contingency and this should all allow for any issues with the project

Chapter 7: **Summary and Conclusions**

Ubiquitous mobile computing and WLAN network usage has increased the importance of an accurate model of human mobility. The list of application domains for such a model is long and ever growing. This report has presented a review of the state of the art in human mobility prediction, location busyness prediction and related areas. Many studies have been done in this area, with various approaches to different forms of the problem being seen in this report. This study focuses on developing a model for predicting the busyness of a location based on WLAN network connection data. This report details all data considerations, including detailing the data source and information. This report also discusses the privacy concerns associated with modelling human movements. Based on the research done during this study, an approach has been detailed and scheduled for the implementation of the remainder of the project. The data will be aggregated by AP and by hour, which will be the training and test data used for training the machine learning algorithms. Three machine learning algorithms will be trained and evaluated : SVMs, Decision Trees, and Naive Bayes. They will be evaluated using hold out testing, and also a number of accuracy metrics ; Classification Accuracy, Precision, Recall, F1 Score and Mean Absolute Error. The most efficient model will be identified and selected. A web application will be built using Django, the most used and most extensive python web framework. D3.js a javascript visualisation library will be used to create the spatial visualisation. The visualisation will be a bubble chart, where each bubble represents an AP and the size and color of the bubble represents the busyness of the node. A user testing session will be setup and testers will fill out a questionnaire to provide feedback on the usability of the application. Feedback will be processed and implemented or detailed as future work. A project work plan has been developed and defined using a gantt chart. An approach has been outlined in detail concerning the development of the web application which will visualise the model. The implementation of the project is scheduled to begin in January 2020 and finish in April 2020.

Chapter 8: Data and Context

This chapter of the report discusses the complete data process of this project. This chapter is divided into three main sections. Firstly there is a short section with some information about the dataset. Then in the second section the data processing steps are described in detail. In the final section the chapter will detail the data considerations, including key decisions and privacy concerns.

The dataset [23] used in this study is described in detail in Chapter 4 of this Report "Data Considerations". The dataset was collected over a 16 month time span from the WLAN network of KTH University. It was initially collected as part of [7] a study titled "Revisiting the modeling of user association patterns in a university wireless network". The study finds that many of the nodes on the network experience time-varying Poisson arrival process, and association distributions can be modeled by two-stage hyper-exponential distributions. This could then be used as a basis for network occupancy models. This study uses machine learning instead to create these models. The dataset contains approximately 95 million records of anonymous users associating with Access Points on the wireless network of KTH Royal Institute of Technology in Stockholm.

Each record contains 3 pieces of data, structured as follows :

timestamp, hashed_user_id, ap_id

Field Name	Description	Format
timestamp	Time at which the association happened.	dd/mm/yyyy mm:hh
hashed_user_id	Hashed Unique ID for user	Length = 40 chars
ap_id	Unique ID for Access Point	'Bldg' + alphanumeric sequence

The required structure of the data for model training is :

building_id, hour, day_of_week, month, busyness_classification

8.1 Data Process

The first stage of the development of this web application is the Data stage. This stage is concerned with all operations, aggregations, manipulations, and decisions relating to the process of transforming this data into the format required for model training. In this section of the report, this process is described. The key issues and decisions involved in the process will be highlighted.

Data Sampling

Due to the large size of the dataset sampling was used throughout the development in order to increase the speed of testing of the model. The dataset contains 95 million records.During development sampling of the dataset was used. Sampling was applied at time of reading the

dataset file by file, so 10% of records in each month's file eg 2015_01 (January 2015) was used. The model did not perform any worse when using sampled data and so sampling was used just to speed up model testing and application testing stages.

Building Extraction

The first step of the data process was to aggregate the APs at a building level. This decision was made to reduce the granularity of nodes in the network from 986 nodes at AP level to 48 at the building level. This aggregation was straightforward as the "ap_id" field in the initial dataset contains values of the format "BLDG1AP10" which refers to the AP in building 1, with id of 10. All AP IDs were prefixed with the building ID, and so to perform this aggregation a regex extraction of the first series of letters followed by a series of numbers was used to aggregate. For example "BLDG1AP10" and "BLDG1AP7" both refer to access points in "BLDG1". This aggregation will make the final visualisation more easily interpreted as a network map of 986 nodes would be confusing and very granular. If the locations of the individual APs were known within the buildings it could be useful to predict a more granular intra-building network map. To extract the "Building ID" field from the "AP ID" field provided in the dataset, regular expression matching (REGEX) was used. A simple regex pattern : "B̂[a-z]*([0-9]*)" which matches a the first series of letters and then the next series of numbers from a string. This extracts "BLDG10" from "BLDG10AP13" for example.

Predictor Variables

The goal of the web application is to allow a user to predict the busyness of a building on the campus at a certain period in time. This is achieved by allowing the user to enter an hour, day of week and month value for which they wish to get a prediction. This functionality requirement was the driver behind the data format requirement for model training. The next step in the data process was to extract the required predictor variables "hour", "day_of_week" and "month" from the "datetime" field in the dataset. These fields are an important part of the model training as they are the three predictor variables our model will be using.

Grouping By Building

The next step in the data process is the counting of connections per building per hour. A grouping function was used to group all connections by building_id and also by hour, and date. This results in a count of how many users connected to each building each hour, for the entire span of the dataset (16 months). EG. :

BLDG1, 01/01/14, 02, 23

This indicates 23 people connected to the Wifi access points in BLDG1 on the first of January 2014 between 1:00 and 2:00 am.

Busyness Scale

The count of users connected to each building is now found. However in the final application, the web app will not tell the user exactly how many users are connected to a node, it will instead give an indication of the busyness ranging from 1 to 5, and this numerical value will correspond to a busyness description which is displayed to the user when they hover the mouse over the buildings bubble. The five descriptions are : "Not Busy"," Moderately Busy", "Busy", "Very Busy", "Extremely Busy". So this process involves going from a raw count of users connected to

nodes for each hour of the entire span of the data collection period, to a rating 1-5 of busyness. To achieve this, all counts were scaled using Min/Max scaling on a per building basis. All the values were scaled to the range of 0 to 1. So if there are 10 people in a building that has a max of 10 this will be very busy, but if a building has 10 people that has a max of 100 this will be classified as not busy. This is beneficial for a number of reasons. The busyness scale will provide a more useful piece of information to the end user instead of just a number of people. This feature also accounts for the fact that some devices may not imply human presence. As discussed in Chapter 4, in [19] and [13], which are both studies using WiFi trace sets to estimate or infer human occupancy, they discuss data considerations relating to the fact that not all WiFi connections indicate human presence. For example if a device is connected for the entire duration of the time period the dataset spans, it is likely that this device is a printer or a server or some other, non human presence indicating device. Using Min/Max scaling does not count these devices , eg. if a building has 5 devices connected to the network for the entire time, its minimum count of 5 will be used in the scaling of connected device counts. These scaled counts of users are then classified on a scale of 1 to 5, or "Not Busy"," Moderately Busy", "Busy", "Very Busy", "Extremely Busy", each representing a classification value from 1 - 5. which will ultimately be used to define a color gradient and size parameters of the bubbles representing nodes in the final visualisation. This visualisation using both the size and colour of each bubble to represent the building's busyness enhances the immediate readability of the visualisation while also accommodating for colour blind users, therefore increasing accessibility as well as the usability of the application.

Outlier Removal

Throughout the time period of data collection, there were certain times when buildings experienced extremely large volumes of users. These extremely large values of user counts, calculated as described above, cause the range of the number of users connected at one time to be very large. When these large min to max ranges were used in the scaling of the data as described above, the scale was skewed and not appropriate. Buildings would never be classified as 3, 4 or 5/5 busyness, because of a single or small few outliers in the dataset. These outliers were removed by setting the "max value for the min max scaling process at a lower value and any value higher than this is set to the this new lower max also. This new max value for the scale is set to the 99th percentile of the buildings values. This seems extremely close to the true max,however upon some exploration of the data it is obvious that is an extremely small number of values for each building that create these extreme max values. This 99th percentile value still maintains a good range of values for all buildings. However it is possible that individual outlier removal on a per building basis may provide a slightly truer representation of the buildings range of user counts. The scale is now effective and values are seen across the scale from 1 to 5.

The data is now in the format required for model training:

building_id, hour, day_of_week, month, busyness_classification

The remainder of this chapter details the decisions that influenced the preprocessing of the data as well as the key points of interest throughout.

8.2 Data Considerations

Month Field

Initially the "month" field was not going to be used for model training. Due to the cyclical nature of a university where people move on timetables and campus is very busy sometimes and also almost empty at other times, it was decided to introduce the month field. In the summer months the university is extremely quiet for example, and so using the month field proved to increase the model prediction accuracy by 10%. This also means a "month" input from the user of the web app is required.

Duplicate User Counts

The dataset used in this study contains records of users associating with the WLAN network, however it does not contain any records of disconnections. This means that if a user for example connects to the network, then connects again 10 minutes later, the user will be counted twice. In section 3 of [7] a number of heuristics are developed based on the protocol and hardware used in the network to estimate the user sessions of the connected users. This was done because the dataset used in the study also has no records of user disconnections, in fact it uses the same dataset as this study. This study also used a technique known as "war-walking" to estimate the radius of WiFi signal transmission from each node in the network. However [7] is focused on modelling user trajectories through a network and also modelling user sessions and patterns, in contrast to this study which is focused on predicting the busyness of a location. Given the level of complexity this session modelling requires, it will not be a part of this study. This is an important thing to consider when using a traceset without disconnections also. However while there are no records of disconnections from the network, there are other techniques to handle duplicate counts of connected users. In this study, the model will be visualised at an hour by hour basis. In the count of users at each node per hour, each unique user is only counted once. This gives a good indication of human presence while also providing a solution to the lack of disconnection records. If a user connects to a node, and connects again 20 minutes later, these two connections still only imply the presence of one human. This means that the size and colour of the bubbles on the bubble map will actually represent the count of unique users that connected to them.

Privacy Concerns

Another concern when doing a study of this nature is that of user privacy. In [7] as discussed above and in [25], user movement is modelled and predicted at an individual level, giving rise to privacy concerns. In [24] it is shown that with high levels of accuracy users can be tracked based on very little data. In this case someone in the university who has very similar mobility patterns which follow their academic timetable could be concerned about the tracking of their movements. However in this study we aggregate the user connections by count, and display this at a building level. In short this avoids all these privacy concerns by focusing on the location level busyness prediction rather than the user trajectory prediction.

Spatial Grid

The dataset used for this project contained an "AP$_{L}ocations.csv"itscolumnsare$:

AP ID, X Coordinate, Y Coordinate

The X and Y coordinates are not latitude and longitude coordinates however. They are reference to

the 0,0 point of a grid created by KTH University. This dataset was initially not released with the relevant latitude and longitude coordinates for the access points. However since the initial release of the data the data provider has made the true latitude and longitude coordinates now . These latitude and longitude coordinates correspond to access point locations on the university, however out visualisation will have the access points aggregated at a building level, and so the bubbles or nodes in the visualisation should represent buildings and not access points. To achieve this a mean latitude and longitude value for all access point coordinates that are in the same building is used as the latitude and longitude coordinate to represent the building.

Data Storage & Data Parsing

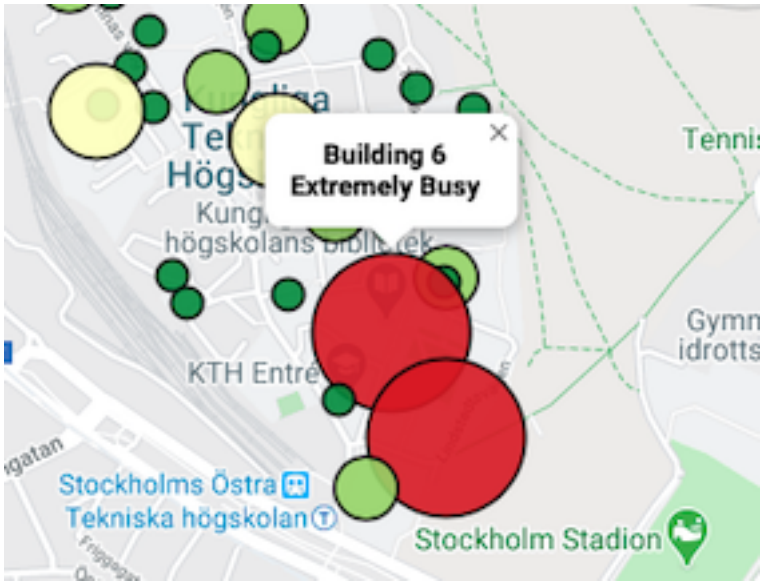
The dataset of records of user associations with APs in the network are stored in 16 .csv files, each representing a month of the 16 month period over which the data was collected. The time period over which the data was collected was from January 2014 to March 2015 inclusive. For model development and training this data is parsed into a pandas dataframe where each row in the datframe represents a record of a user association, and each column represents its respective value from the record, as described at the beginning of the chapter. Once the data is in a pandas dataframe it is easily accessible and all operations can be performed on it using python. The model is easily accessed by the web application which is also python based (Flask). This means all in one process data can be loaded, parsed, cleaned, manipulated, modeled and visualised.

Chapter 9: Detailed Design and Implementation

This section will describe the design and implementation of the web application developed as part of this study. Figure 1 below contains a screenshot of the web application. Figure 2 below contains a screenshot of the tool-tip functionality that is part of the web application. This tool-tip is described in detail below.

Figure 1 : Web Application

Figure 2 : Tool-Tip



Functionality of Base Web Application

The web application's main functionality is to provide a map-based visualisation of the university campus on which visualises the busyness of the various university buildings. This specific functionality means that the web application does not have any requirement for multiple pages, multiple forms or multiple use cases. This very specific use case and functionality informed the design of the web application, the design is minimalist, offering the user an easy to use user interface which does not require any time to learn. The web application is a single page application.

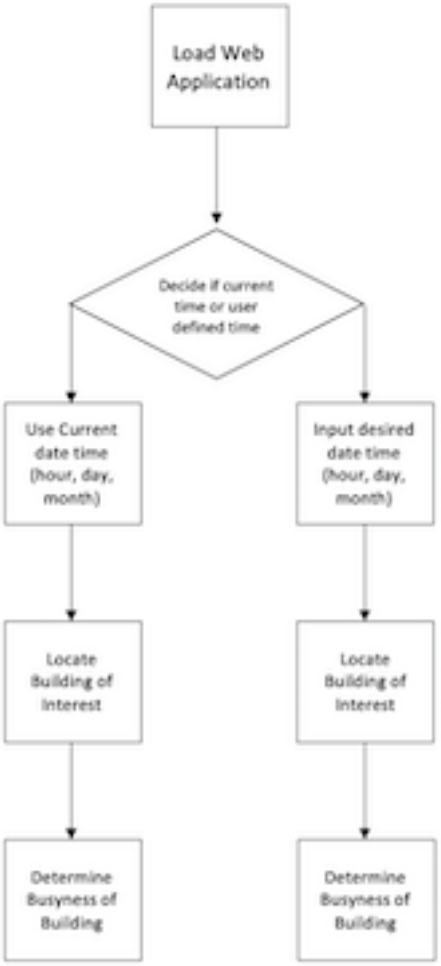
Figure 1 above is a screenshot of the web application. At the top of the page the title of the web application is displayed: "KTH University - How Busy?". The web application contains a simple three field form at the top of the page. The three fields are Hour, Day and Month. When the user loads the web application for the first time, these three fields default their values to the current Hour, Day and Month at the time the user loads the web application.

The map displayed as the main component of the page displays the university campus using the Google Maps API. The requirement for this visualisation is that it spatially visualises the busyness of building locations on the university campus. The visualisation is designed to achieve this is a bubble map. This is a type of visualisation where there are nodes or bubbles overlaid on a map, and where the size of the node or bubble represents an attribute of the node. The map displays a bubble for each building on the campus, the bubbles are coloured and sized proportional to their busyness. If the user has just loaded the web application for the first time this map will display the predicted busyness for each building at the current Hour, Day and Month when the user loads the web application. If the user inputs a desired Hour, Day and Month the map will be rendered to display the prediction for the given user input.

The bubble map provides all the in built capabilities of google maps, allowing the user to zoom, use satellite view and go full screen. If the user wishes to check the busyness of a building, there is a tool-tip which makes this very easy. This tool-tip is pictured in Figure 2 above. If the user hovers over a building's bubble, an info window is displayed as a tool-tip. This tool-tip shows the building name and the building's busyness description. If the user moves the mouse off the bubble, the tooltip dissapears. If the user clicks on a particular building's bubble the map pans to that building aas its center and zooms in.

This functionality is displayed in a UX diagram below in Figure 3:

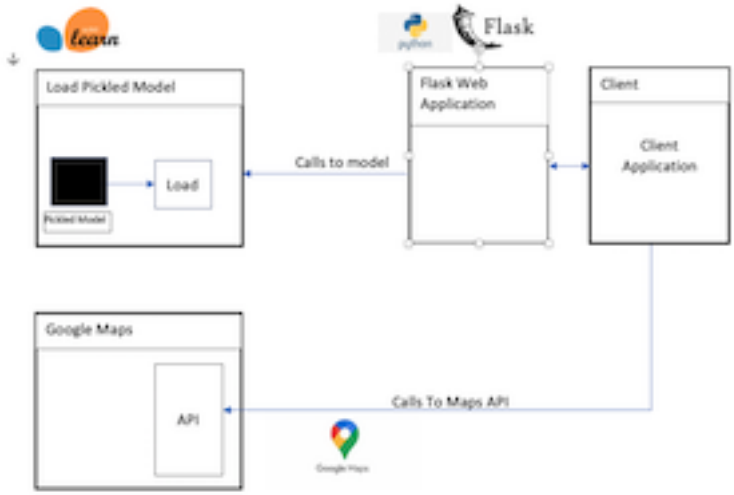
Figure 3 : User Flow Diagram :



Implementation of Web Application

This section will describe the implementation of the web application in detail. It will discuss the tools used to develop the web applicaiton and how they interact to produce the final product. Figure 3 below is an architecture diagram which gives an overview of the web application structure.

Figure 3 : Architecture Diagram



The web application was developed using the flask python library [1]. Flask is a lightweight library that is very suitable for a project of a narrow scope. As described above in Chapter 8: Data Context, all of the data preprocessing, manipulation and model training was done in python also. The python module pickle was sued to save the model after creation. This made the integration of the model with the web app extremely convenient. When the application is run, it loads the pickled model.

When the application receives a request, it calls the view function for that URL, this is the basis of flask web development, each URL request calls the view function associated with that URL. In this application the view function sets the hour, day and month values equal to the current hour, day and month if it is a GET request. In the case of a POST request, the hour, day and month values are set to be the values passed by the POST request triggered when the user pressed the submit button in the user input form on the web application front end. Next the flask view function requests a prediction from the model for the given hour, day and month and stores the response prediction. The view function then calls the render function for the HTML template, passing the predictions and also passing the form to be used. Flask provides support for web forms. A flask web form with three input fields "Hour", "Day", and "Month", and a submit button is passed to the HTML page. The HTML page conatins all the html code to renders the title, and web form and has a large section of the page set to render the map, this section is a large div with id set to map. JavaScript is then added which calls the google maps API and applies it to this div.

The javascript code creates a google maps map object and inserts it into the html div with id "map". the code adds a bubble to the map for each building, and the flask web application has passed a busyness prediction for each of these buildings. In the case of this application each node on the map will represent a building in the university campus. Each node or building is centred on the mean of the latitudes and longitudes of the WiFi access points associated with each building. This is due to the fact that only latitude and longitude coordinates for individual access points are provided, and not building specific coordinates. The busyness prediction passed for each building will define the colour and size of the building's bubble. Using size and colour here can enhance the usability and accessibility of the visualisation and make it easier for the user to quickly identify the nodes on the map. In [26] a study on the use of colour in data visualisation, it is discussed that colour is a very effective retinal variable. The study mentions Color Brewer [2] which offers some very useful colour schemes. For the purpose of this visualisation a divergent colour scheme generated by Colour Brewer which has 5 colours form red to green is used. This scheme offers an intuitive

[1] http://flask.palletsprojects.com/en/1.1.x/
[2] https://colorbrewer2.org/type=sequentialscheme=BuGnn=3

qualitative visual measure in the visualisation. Small nodes will be green, indicating low busyness, while large nodes will be red, indicating high busyness. The busyness is used as an index to get the relevant colour from the array. the colours used go from green to red and their hex codes are : "57f542","FFFF00","ffbe42","FF8C00" and "FF0000". The nodes are also set to have an opacity of 0.9, this allows the user to see under the node also which can allow the user to see some relevant information from the map underneath, especially when the nodes are larger and cover more map space. The bubbles also have a black border which creates a contrast against the map background making the bubbles stand out to the user more easily. Each node is also assigned a busyness description which one of five values, again depending on its predicted busyness rating. These values are : "Not Busy","Moderately Busy","Busy","Very Busy", "Extremely Busy". The google maps map object is set with it's initial center to be a central latitude and longitude coordinate in the university campus, and the zoom is set so the map is relatively zoomed in on the campus. Each node also has a google maps "info window" which is used to provide the tool-tip functionality. This tool-tip is a small window which displays the building name and the busyness description. The tool tip is displayed using a "mouse over" JavaScript listener event so when the user mouses over the node, its respective tool-tip is displayed. Similarly the tool-tip disappears when the user moves their mouse away from the building, this is done using a "mouse out" JavaScipt listener event. Each node also has a "click" JavaScript listener event. If a user clicks on a node, the map is centered on that nodes coordinates and the zoom is also set to a slightly increased value. This is to provide the user with an easy way to focus their attention on the building of interest.

Chapter 10: **Evaluation**

In this section of the report a detailed evaluation of the study will be carried out. This evaluation is divided into two main sections. The first section will detail the evaluation of the machine learning model developed for this study. This evaluation will look at a number of evaluation metrics : Prediction Accuracy, precision, recall, f1 score, and also the mean squared error. The next section will detail the evaluation of the web application implemented as part of this study. This evaluation was conducted by means of an ASQ (After Scenario Questionnaire) questionnaire distributed to test subjects. The survey and its results will be detailed here.

Model Evaluation

In this study the dataset is used as training and test data for the machine learning algorithm that makes predictions that are used in the web application. The model is at the core of the functionality of the web app and is the basis for its development. The machine learning algorithm is the key feature in this study used as a solution for location busyness prediction. The three predictor fields it uses are the hour, day of the week and month of the year. The dependent variable or the variable that it predicts is busyness. For each building id in the dataset, given a time and date, the model predicts a busyness rating from 1 - 5. The busyness value corresponds to a level of busyness : 1 = Not Busy, 2 = Moderately Busy, 3 = Busy, 4 = Very Busy and 5 = Extremely Busy. The python module scikit-learn is used in this study. It is a free machine learning python module which provides extensive support for machine learning algorithms including training, testing and evaluation.

To choose the right model to use the data and the task of prediction must be assessed. The data used for the training of the data is labelled data. Meaning the values for busyness have already been found. This is described above in the data and context section. The counts of users per building per hour were found and scaled, resulting in busyness ratings which is why this is "labelled data". A supervised learning approach can be taken with this labelled data. The task of predicting a value from 1-5 is essentially a classification task, and so models which perform classification are suitable here. The dataset is very large with 95 million records. The three classifiers chosen to be tested were Decision Tree Classifier, K Nearest Neighbours Classifier and Naive Bayes. Each of these models was evaluated before the final model, that was integrated into the web application, was chosen.

Each model that was trained was tested under a number of different performance metrics. The model used an 80 / 20 split for training and test data. this split was decided by evaluating the performance metrics of the model with different split values. Starting at 10 / 90 and incrementing by 10 each iteration until 90 / 10. Although the models performed very similarly with various values for the split the 80 / 20 split performed the best. In the case of this data set 10% of the dataset is still 9.5 million records. The very large size of the dataset means the model has plenty of data for both training and validation. The performance metrics used to asses the models' performances are : Classification Accuracy, Precision, Recall, F1 Score and Mean Squared Error. These metricss are widely used in the evaluation of machine learning algorithms. In studies [**6068488**], [**article**] and [**10.1145/1568199.1568210**] classification accuracy and/ or RMSE are used as evaluation metrics. The classification accuracy is the rate of correct classifications. RMSE is the Root Mean Squared Error. Root Mean Square Error is the standard deviation of the residuals(prediction errors). Residuals are a measure of how far from the regression line data points are (How incorrect they are); RMSE provides a measure of how spread out these residuals are. RMSE tells you how concentrated the data is around the line of best fit. Initially it was planned to use Mean Absolute

Error, but Mean Square error differs to this by making larger errors add more to the score. This is useful in this scenario as it is much worse to predict that a building which is 1/5 busyness is 3/5 busyness than it is to predict that it is 2/5 busyness. A larger error here makes the models output significantly less useful to the user. This is why Mean Squared Error was selected instead. Precision calculates the proportion of positive predictions, those are actually correct. Recall calculates the proportion of actual positives that were identified correctly. F1 score provides an average measure of recall and precision. These metrics are widely used to examine the predictive performance of a classifier. The following table details the values that each model got under assessment of each metric :

Model Name	Prediction Accuracy	Precision	Recall	F1 Score	Mean Squared error
Decision Trees	94%	0.87	0.746	0.742	0.122
KNN	92.5%	0.863	0.735	0.736	0.1641
Naive Bayes	88.15%	0.778	0.882	0.827	0.507

It is clear from the above table that the Decision Tree Classifier outperformed all other classifiers. It has the highest classification accuracy at 94%. The Decision Tree Classifier also scored the highest for precision meaning it has the highest percentage of true positive classification. The Decision Tree Classifier also has significantly lower Mean Squared Error. Decision Tree Classifier is the optimal model and will be used for classification for the remainder of this study.

Web Application Evaluation

The Web Application built as part of this study is an application that allows the user to predict the busyness of the buildings on the university campus of KTH Royal Institute of Technology in Stockholm Sweden. The web application was built with both usability and accessibility in mind. The bubbles on the bubble map each represent a building on the campus. Each bubble's colour represents its busyness rating. Each bubbles size also represents its busyness and so the application is accessible even to colour blind people. There is also a very useful tool-tip when the user hovers over a bubble on the map. This tells the user the name of the building as well as its busyness description from Not Busy to Extremely Busy. These features enhance the usability and accessibility of the application. The user input form is also minimal and placed at the top of the page making it obvious to the user how the web application works. The application uses a minimalist design along with the extremely user friendly experience provided by the Google Maps API to provide a straightforward user experience.

The web application is assessed by means of an ASQ or After Scenario Questionnaire. This is a widely used questionnaire for usability evaluation, due to it's short time taken to do, as well as it's proven reliability. This questionnaire is discussed in detail in [27] This is a standard questionnaire used to evaluate usability of a system. It was created by IBM and is widely used. The questionnaire is filled out by test users immediately after performing a process. It is a very simple questionnaire with three questions :

1. Overall, I am satisfied with the ease of completing the tasks in this scenario.
2. Overall, I am satisfied with the amount of time it took to complete the tasks in this scenario.
3. Overall, I am satisfied with the support information (online-line help, messages, documentation) when completing the tasks.

Each question is answered with a number from 1 to 7, where 1 means strongly agree and 7 means strongly disagree.

The participants for the survey were 50% university students and 50% members of the general public.The university students were selected as participants because the application is in the domain of a university. The students would have a general understand of campus life and of the requirement for busyness prediction. Ideally these students would have been students from the university itself KTH Royal Institute of Technology. However this was not possible for the purpose of this study. The other 50% were members of the public. This was to examine the usability of the application amongst people with no knowledge of campus or university life. Each test user was given a defined scenario.The complete test scenario and questionnaire can be found in the appendix of this report. The task the user was to complete was to identify the busiest building(s) on the campus. The questionnaire was filled out by 20 test users and the responses to each question in the questionnaire are detailed below. The survey included a GDPR statement which is detailed below in this section.

Question 1 Of the 20 responses to this question 17 users or 85% of users responded 1, while the remaining 3 users or 20% of users responded 2. These responses give a good indication that all users were able to complete the task, and almost all of them were able to complete the task with complete ease.

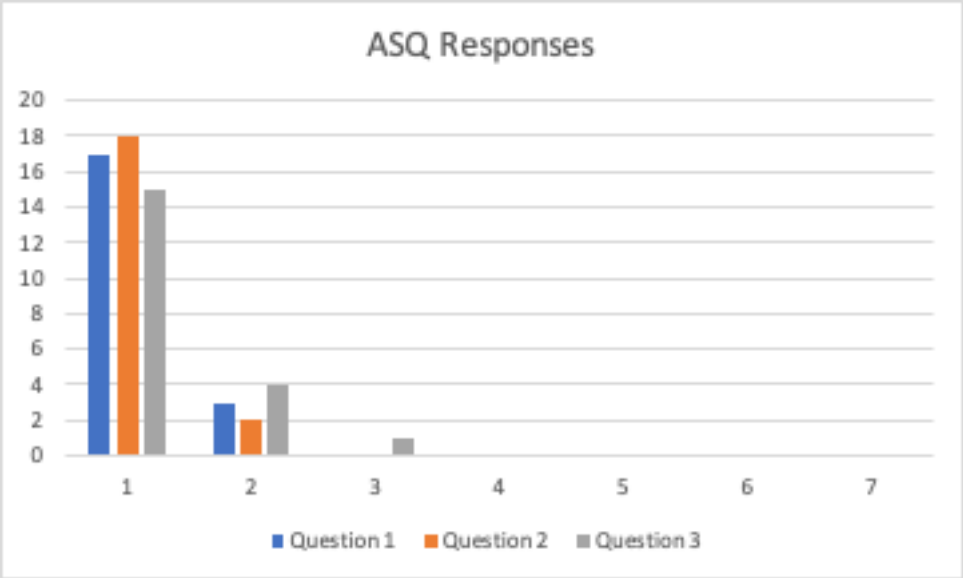
Question 2 Of the 20 responses to this question 18 users or 90% of the users responded 1, while the remaining 2 users responded 2. These responses give a good indication that all users were able to complete the task within a reasonable amount of time, and almost all users saying they were completely happy with the time it took to complete the task.

Question 3 Of the 20 responses to this question 15 users or 75% of the users responded 1, 4 users or 20% of the users responded 2 and 1 user or 5% of the users responded 3. These responses give a good indication that all users felt the application provided support to them. This is likely through the tool-tip.

These responses overall show that all users were very satisfied with the usability of the web application. The minimalist design combined with the colour and size bubble design and the tool-tip on hover all provided the user with an easy to use user interface. This can be seen in the user responses. In general the user feedback was very positive, some informal feedback was received also. A number of users commented on the tool-tip, saying it worked very well and exactly how they expected. The on click feature where the map zooms to the selected building and centres on it also received a few comments. These features seem to have achieved the exact support they were intended to.

The results are summarised in the following clustered bar chart in Figure 1:

Figure 1 : ASQ Response Data :



The collection of the responses to the questionnaire was GDPR compliant and contained privacy policy statement at the beginning of the questionnaire informing the users about the information collected etc. The statement can be found in the complete copy of the questionnaire in the appendix of this report

Chapter 11: Conclusions & Future Work

This study will include a report conclusion, as ell as a discussion around any potential future work in this area.

To re-state the problem this study addressed :

"Given a dataset containing records of user association with nodes on a WLAN network, create an algorithm to predict the busy-ness of an Access Point (AP) on the WLAN network. In addition to this, create a web-based application that spatially represents the network and the network users over time."

This study has completed a thorough literature review of the state of the art in the field of human mobility analysis and mobility prediction. This study "Wireless Network connections as a Proxy for Human Presence" has shown that wireless network connections are indeed a very good proxy for human presence, in this case allowing a model to be created which can perform location busyness prediction with a classification accuracy of 94%. A system was created using a number of tools to create the required model, and also to visualise it in a web application which spatially visualises the model. The model which performed the best was a Decision Tree Classifier. It has been evaluated by a number of performance metrics: Classification Accuracy, Precision, Recall F1Score, and Mean Squared Error. The models results are displayed below :

Model Name	Prediction Accuracy	Precision	Recall	F1 Score	Mean Squared error
Decision Trees	94%	0.87	0.746	0.742	0.122

A web application was built using flask, a python web framework. The Google Maps API was used to create the spatial visualisation of the campus. The visualisation is a bubble map, where each building on the map is represented by a bubble. The predicted busyness for each building is represented by the size and colour of the building's bubble. The web application was evaluated in terms of usability through means of the widely used IBM's After Scenario Questionnaire (ASQ). The results of the questionnaire were extremely positive, indicating the success of the design and the features to create a usable and intuitive user experience.

Evaluation Of This Study

This subsection discusses the project's success in relation to the overall objectives as defined in Chapter 1: Project Specification. In terms of the core objectives all objectives were completed. The advanced section of the project has also been completed. 100% of the project specification has been completed, both core and advanced. The end result of this project can be described accurately by the opening paragraph of the project specification :

"Given a dataset containing records of user association with nodes on a WLAN network, create an algorithm to predict the busy-ness of an Access Point (AP) on the WLAN network. In addition to this, create a web-based application that spatially represents the network and the network users over time."

The state of the art in human mobility analysis is reviewed in this paper. Most of this is reviewed and discussed in Chapter 3, but several other references to relevant state of the art are also

described throughout. The data has been fully described in Chapter 4: Data Considerations and in Chapter 8: Data and Context. Suitable features for predicting location busyness were identified in the dataset. A Decision Tree Classification Model has been developed to predict the busy-ness of a location. This model is described in detail and this approach has been thoroughly evaluated in this chapter Chapter 11: Evaluation. an interactive web based application has been developed to display the model and its predictions.

Although this project achieved all of it's objectives, there are many variations to the approach that could have been taken. This study has outlined a number of areas in which there is potential for further work, which could potentially improve the approach taken. The field of human mobility prediction is huge, and involves many interesting concepts. Potential for future work in this area, specifically relating to this project, is detailed below in the Future Work section.

Future Work

The field of human mobility prediction and analysis is a huge field with many different application domains such as urban planning, epidemiology, network resource allocation [1], transport network design and disaster relief [2]. There are also many different approaches which have been taken to the problem of human mobility prediction. These various approaches are discussed in detail in Chapter 3.2 "Different Predictors".

This study shed light on a number of different ways in which the are could be further developed, and there is much potential for future work.

One potential piece of future work would be to also incorporate a dataset of weather related data. For obvious reasons weather can have an impact on human mobility and so could provide significant benefit to the model's predictions. Weather data has been used in many machine learning models like in [28] which uses weather data to predict air traffic delays. Also in [29] where weather data is used to predict renewable energy production.

Another potential piece of future work in this study arises from the fact that the dataset used only contains records of user associations with the network. It does not include any disassociation records. This means the length of a user's session on the network can never be known. However it can be estimated. This session estimation is explored in [7]. This could potentially be very useful if the application is to predict demand for network resources or to more exactly track the movements of individuals.

This study focuses on location busyness prediction in a university campus, however it could also be useful to know the busyness prediction for the routes between buildings. This could be done by looking at where a user previously appeared when they are seen at a node. This allows you to know the route they probably took to get between the two buildings. At a high level this would allow you to also imply route busyness prediction which could be an extremely useful application. This concept is explored in [18].

The field of human mobility is ever growing and changing, and there will constantly be a demand for new research in the field and for the development of new technologies and systems.

This study can conclude that wireless network connections can be used as training data for a model to predict the busyness of locations.

Appendix

ASQ (After Scenario Questionnaire)

Below is a complete copy of the ASQ (After Scenario Questionnaire) which was filled out by 20 participants.

Privacy Statement:

1. The only data collected is your responses to the following 3 questions on the usability of the web application.

2. This data is being collected as part of an academic research project. The data will not be shared with any third parties.

3. The data will be used as an indication of the usability of the web application. This is the only use for this data.

4. This data will only be retained until this study has been completed. The completion data for this project is May 21st 2020.

5. The data collected here is completely secure and will not be shared with nay third parties. It is anonymous and no personal information will ever be at risk.

6. As a participant you have the right to access view or edit your response at any point in the future. You also have the right to request the deletion of your response data.

Test Background : This web application provides a prediction for how busy the buildings in the university campus of KTH Royal Institute of Technology in Sweden is. For the purpose of this test, the user scenario is that the user wants to identify the busyest building(s) on the campus on Monday the 4th of May at 19.30 hours.

After the completion of this test scenario you will fill out the questionnaire below. There are three questions, each which is answered with a value from 1 to 7, where 1 means strongly agree, and 7 means strongly disagree. Select answer by ticking the respective box.

Test Case:

Identify the busiest building(s) on the campus on Monday the 4th of May at 19.30 hours

Task Steps:

	Step	Step Data
1	Go to site	URL
2	Enter Hour	19.30
3	Enter Day	Monday
4	Enter Month	May
5	Identify busiest building(s)	

After Scenario Quesitonnaire:

1) Overall, I am satisfied with the ease of completing the tasks in this scenario

1	2	3	4	5	6	7

2) Overall, I am satisfied with the amount of time it took to complete the tasks in this scenario

1	2	3	4	5	6	7

3) Overall, I am satisfied with the support information (online-line help, messages, documentation) when completing the tasks

1	2	3	4	5	6	7

Acknowledgements

I would like to thank Dr Gavin McArdle of the UCD School of Computer Science, my project supervisor, who helped and guided me through this project.

Bibliography

1. Ganji, F. *et al.* Greening campus WLANs: Energy-relevant usage and mobility patterns. *Computer Networks* **78,** 164–181 (2015).
2. Song, X., Zhang, Q., Sekimoto, Y. & Shibasaki, R. *Prediction of human emergency behavior and their mobility following large-scale disaster* in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (2014), 5–14.
3. Liu, Y. *et al.* Intelligent bus routing with heterogeneous human mobility patterns. *Knowledge and Information Systems* **50,** 383–415 (2017).
4. Depatla, S., Muralidharan, A. & Mostofi, Y. Occupancy estimation using only WiFi power measurements. *IEEE Journal on Selected Areas in Communications* **33,** 1381–1393 (2015).
5. Pajevic, L., Fodor, V. & Karlsson, G. *Predicting the Users' Next Location From WLAN Mobility Data* in *2018 IEEE International Symposium on Local and Metropolitan Area Networks (LANMAN)* (2018), 61–66.
6. Wang, W., Chen, J. & Hong, T. Occupancy prediction through machine learning and data fusion of environmental sensing and Wi-Fi sensing in buildings. *Automation in Construction* **94,** 233–243 (2018).
7. Pajevic, L., Fodor, V. & Karlsson, G. *Revisiting the modeling of user association patterns in a university wireless network* in *2018 IEEE Wireless Communications and Networking Conference (WCNC)* (2018), 1–6.
8. Jahromi, K. K., Zignani, M., Gaito, S. & Rossi, G. P. Simulating human mobility patterns in urban areas. *Simulation Modelling Practice and Theory* **62,** 137–156 (2016).
9. Gao, H., Tang, J. & Liu, H. *Mobile location prediction in spatio-temporal context* in *Nokia mobile data challenge workshop* **41** (2012), 1–4.
10. Vintan, L., Gellert, A., Petzold, J. & Ungerer, T. Person movement prediction using neural networks (2006).
11. Toch, E., Lerner, B., Ben-Zion, E. & Ben-Gal, I. Analyzing large-scale human mobility data: a survey of machine learning methods and applications. *Knowledge and Information Systems* **58,** 501–523 (2019).
12. Anagnostopoulos, T., Anagnostopoulos, C., Hadjiefthymiades, S., Kyriakakos, M. & Kalousis, A. *Predicting the location of mobile users: a machine learning approach* in *Proceedings of the 2009 international conference on Pervasive services* (2009), 65–72.
13. Wang, Z., Hong, T., Piette, M. A. & Pritoni, M. Inferring occupant counts from Wi-Fi data in buildings through machine learning. *Building and Environment* **158,** 281–294 (2019).
14. Song, C., Qu, Z., Blumm, N. & Barabási, A.-L. Limits of predictability in human mobility. *Science* **327,** 1018–1021 (2010).
15. Smith, G., Wieser, R., Goulding, J. & Barrack, D. *A refined limit on the predictability of human mobility* in *2014 IEEE International Conference on Pervasive Computing and Communications (PerCom)* (2014), 88–94.
16. Yoo, S.-e. *et al. PGS: Parking Guidance System based on wireless sensor network* in *2008 3rd International Symposium on Wireless Pervasive Computing* (2008), 218–222.
17. Toole, J. L., Ulm, M., González, M. C. & Bauer, D. *Inferring land use from mobile phone activity* in *Proceedings of the ACM SIGKDD international workshop on urban computing* (2012), 1–8.

18. Akagi, Y., Nishimura, T., Kurashima, T. & Toda, H. *A Fast and Accurate Method for Estimating People Flow from Spatiotemporal Population Data.* in *IJCAI* (2018), 3293–3300.

19. Vaccari, A., Samouhos, S., Glicksman, L. & Ratti, C. *MIT enernet: Correlating WiFi activity to human occupancy* in *Proceedings of Healthy Buildings* (2009).

20. Kalikova, J. & Krcal, J. *People counting by means of wi-fi* in *2017 Smart City Symposium Prague (SCSP)* (2017), 1–3.

21. Naga, A. *Estimating the number of people in an area: Using Bluetooth and WiFi signals* 2019.

22. Ouf, M. M., Issa, M. H., Azzouz, A. & Sadick, A.-M. Effectiveness of using WiFi technologies to detect and predict building occupancy. *Sustainable buildings* **2,** 1–10 (2017).

23. Pajevic, L., Karlsson, G. & Fodor, V. *CRAWDAD dataset kth/campus (v. 2019-07-01)* Downloaded from https://crawdad.org/kth/campus/20190701. July 2019.

24. De Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M. & Blondel, V. D. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports* **3,** 1376 (2013).

25. Lin, Y., Huang-Fu, C. & Alrajeh, N. Predicting Human Movement Based on Telecom's Handoff in Mobile Networks. *IEEE Transactions on Mobile Computing* **12,** 1236–1241 (2013).

26. Lee, S., Sips, M. & Seidel, H. Perceptually Driven Visibility Optimization for Categorical Data Visualization. *IEEE Transactions on Visualization and Computer Graphics* **19,** 1746–1757 (2013).

27. Lewis, J. R. Psychometric Evaluation of an After-Scenario Questionnaire for Computer Usability Studies: The ASQ. *SIGCHI Bull.* **23,** 78–81. ISSN: 0736-6906. https://doi.org/10.1145/122672.122692 (Jan. 1991).

28. Choi, S., Kim, Y. J., Briceno, S. & Mavris, D. *Prediction of weather-induced airline delays based on machine learning algorithms* in *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)* (2016), 1–6.

29. Kim, S.-G., Jung, J.-Y. & Sim, M. K. A two-step approach to solar power generation prediction based on weather data using machine learning. *Sustainability* **11,** 1501 (2019).