Understanding Group Human Mobility Behaviour using a University Campus Data Set

By

James Little

Department of Computer Science
University of York

In Partial Fulfilment of the Requirements
for the Bachelor of Science

May 2021

Supervisor
Iain Bate

Table of Contents

Executive Summary

From a basic review of human mobility, it is clear that there is a lack of research linking the classification of individuals/groups within a system; this paper seeks to achieve this. Being able to understand group behaviours, create links between unlabelled data and group behaviours is a significant technique. In this case using a university campus dataset, provides an approach that could be utilised to aid the design of commercial spaces to maximise utilisation of spaces, design evacuation routes, and target group behaviours and movement patterns of classified groups of people.

Using an unlabelled dataset based on data from the connection logs for all the WiFi access points on the KTH University Campus in Sweden the study used a number of clustering algorithms to classify and analyse the data, including a technique to add value to WiFi connection logs in order to aid the classification.

The research tested a range of unsupervised clustering schemes and utilised data processing techniques to aid the classification and analysis of the dataset, including adapting the dataset to summarise user movements through the system[1]. The implementation used three common machine learning pre-processing techniques: correlation matrix; PCA dimension reduction; and, MinMax normalisation. To evaluate the results using this method, the approach allowed for individuals to be characterised into groups against a number of differentiating criteria in effect creating group ground truths.

The analysis of this research proved that from the tested techniques, Gaussian mixture clustering provided the most consistent scheme for classification of the unlabelled dataset. The correlation matrix approach was ineffective when working with the limited number of parameters within the processed dataset. However, using normalisation and PCA dimension reduction increased the accuracy of model classification. This approach, utilising the feature creation scheme, pre-processing techniques and gaussian mixture clustering has potential to be utilised with other human mobility systems such as shopping centres, transport hubs and government buildings.

The challenge of this type of research is the limitation of unlabelled datasets and WiFi access logs, being able to implement summarising features significantly boosted the interpretability of the data and consequently made classification not only possible but accurate. This paper also reflects on issues related to social and ethical implications of using data which is characterised as sensitive and protected.

The research has been bounded and in part limited by the datasets, the focus is on the approach to the analysis utilising readily available data, the research

[1] An area of contained human movement, in this case the Campus.

has however demonstrated/proved an analysis framework that could be applied to datasets of varying scenarios using the same data acquisition techniques.

Introduction

Background

How a human interacts and moves within a system has always been a challenge for many fields of study, but predicting and analysing future movements has previously been considered near impossible due to the challenges of data acquisition, and computational complexity. However, the increase in mobile phone usage with research suggesting near 90% ownership [1], and advances in data collection and computational techniques has seen progress of research in this field including applications in Physics [2] and Geography [3].

The study of how individual/groups of humans move within a network or system [4] is considered as the domain of human mobility research which spans many subjects including data privacy, data processing and analysis, and data collection, all of these are covered to some degree in this research. In assessing human mobility analysis, this research focuses specifically on using Spatiotemporal data and analysing it to both categorise and provide further insight. Current human mobility analysis is focused on the prediction of human behaviour given past movements; these studies have found that although human behaviour has a desire for change and spontaneity, our daily mobility is characterised by a deep-rooted regularity [5]. This has led to the creation of human mobility models that are both sufficiently accurate and adaptable [5], [6], [7].

Although a key idea within human mobility analysis is the ability to predict future movements, the research in this area encompasses much more. This area of research has inherent privacy concerns that raise ethical issues, and there is related research into this area has looked to provide solutions to stop malpractice [8] and misutilisation. The opportunity for large-scale human mobility information and trend analysis has led to an increase in practical interest from non-academic research and commercial applications. With this interest, more human mobility datasets have been gathered allowing for questions to be answered on the problem once thought immensely difficult. One such topic is the spreading and tracking of infectious biological diseases, which in recent years has had a research spike due to the 2019 pandemic [9]. Although COVID-19 has been at the forefront of research, Ebola and Malaria have been a part of other such human mobility studies [10], [11]. On the topic of public health, research has been conducted into Identifying public health trends and classifying an individual’s weight and health [12].

Looking to more Commercial/Governmental applications of human mobility analysis, the techniques in this research are applicable to a range of urban settings where the analysis of data can provide critical insight for those involved in urban space optimisation. The techniques can also be retrospectively applied to alleviate legacy city planning. Research has investigated such topics as mass movement trends of people for Intercity travel [], public transport design through mobility trends [13], metropolitan hotspot analysis [14] and practical taxi-cab allocation to account for travel demand [15]. It would be interesting to see how this approach could be utilised to modify population flows through congestion points in urban settings.

Predicting human mobility is a large part of current related research, but there is a strong argument for other sub-topics of study being just as important. Two further areas encompassed within the wider topic address the problem of data acquisition and project scale. Project scale refers to the system on which analysis is to be conducted, due to a large influx of cellular data in the early 2010s larger studies have been very popular with Metropolitan areas having a large amount of analysis conducted. In some cases, even larger studies have been conducted with whole countries participating [16]. An area that has received less interest from the academic community has been that of indoor location analysis, this has been for multiple reasons but mainly due to the increased complexity to retrieve data. Such studies can have increased gain with an individual's mobility having a greater value to it [17] [18] than a group's mobility. In this respect mobility value refers to the possible educational and commercial applications that the data can provide.

In addition, this had a direct link to how data is acquired within human mobility analysis which has been the catalyst for research to help compare methodologies to find optimal techniques. Within the community, numerous techniques and technologies have been used to acquire mobility data, enabled by the uptake of; GPS [19], Cellular Data [20], RFID [21] and Wi-Fi [22]. All of which provide key differences to each other and have scenarios that they work best in.

Human mobility analysis has proven over the last 13 years since the first large scale human mobility model [6], that the techniques and technologies created through its research have been very effective. This has provided motivation to progress further into the research space whilst being able to apply a unique approach to this already saturated field of study.

Literature Review

Human mobility is a multi-faceted area of study, this literature review has been split into: the prediction of future human mobility; data acquisition techniques; and privacy considerations.

Predicting Future Human Mobility

This is a key focus of related research into human mobility, with by far the largest amount of research conducted within the last 10 years. Most papers link to model creation in some way to predict a future location and then build upon this by undertaking research into other facets of study such as privacy implications and data acquisition. Within these models, many systems are outlined with metropolitan studies being the most popular.

Human mobility modelling first started with a preliminary paper by D. Brockmann, L. Hufnagel & T. Geisel [23], this research conducted analysis on the subject which at that time was widely known as human travelling statistics. It used the circulation of banknotes in the USA to assess human movement. The study found that banks notes, conform to a scale-free random walk (Levy Walk) for which an assumption was made that human mobility would act similarly. The study went further by modelling the random walk to produce a surprisingly accurate model, thus confirming their research but due to the infancy of the model, questions were raised on the assumption of correlation between mobility trajectories. The research provided a good introduction for the area whilst also leaving much to be desired, as I argue that although there are similarities between the two human trajectories offer much more complexity than that of a tracked banknote. Marta C Gonzales et al [6] looked to build upon the 2006 paper. This unlike the paper before used a human spatiotemporal dataset recorded over 6 months. The team found that the results contrasted with the previous study noting that “human trajectories show a high degree of temporal and spatial regularity” [Ref] which differed from the more random assessment made previously. Both studies are still important as they show a problem that current research is still trying to solve, that although human mobility is regular and predictable there are always cases of temporal and spatial irregularity.

More recently Eunjoon Cho et al [24], looked to predict movement from two factors; spatiotemporal data and social media check-ins (Twitter, Facebook etc.). They hoped to explain the amount of an individual’s mobility that can be gleaned from social relationships and their periodic behaviour. This study found that social relationships can explain 30% of our mobility data whilst our ‘day to day’ movements explaining the rest. This research provides the needed proof to explain anomalies that are consistently present within data as well as indicating how sporadic mobility can be within social scenarios. Recognising human mobility’s close relation big data, K Zhao et al [25] looked to offer a one size fits all framework to human mobility modelling from a data mining point of view. This led to a soup of machine learning and data mining methods which were then applied to current human mobility datasets and studies. The addition of methods; data cleaning/pre-processing and mining of patterns within current machine learning models, lead to the discovery of the many benefits that a data mining approach can provide as well as enhancing the framework to easily use and build upon their research.

Mobility models do not always lead to spatiotemporal predictions but are able to use mobility data to categorise agents or locations within the system based on behaviours. One of the best examples of such research used geospatial data to characterise complicated urban development known as mixed-use [26]. This mixed-use development district was introduced to blend residential, commercial, institutional premises together into an integrated real estate type. This functionality caused complications during the identification and categorisation of these districts. The study used a probabilistic, Bayes theorem-based model that gave an 85% accuracy when referencing urban survey data. This area of mobility analysis has very little research recently, such studies could provide valuable data to help city planning and governments with policymaking.

In contrast with the recent uptake in social media applications, the process of using geotagged locations has been able to provide large scale high resolution public data. One such study that has used this to model human mobility, has used geotagged tweets posted in Australia from Twitter. Raja Jurdak et al [27] aimed to justify metropolitan trends whilst also proving that such data can be used for both the tracking and predicting of human mobility. This was done through a probabilistic entropy based prediction model that proved the adaptability of this data. The benefits of such data being used for mobility research is its innate availability and high resolution, but from this comes a large re-identification privacy issue.

More recently the population migration analysis has dominated research due to the COVID-19 pandemic, Jayson S. Jin et al [28] investigated national aggregate population flows around the epidemic epicentre from when mass transmission was recorded. The study had two goals: one to measure the effectiveness of quarantine and the other most importantly was to create a model to accurately predict province hotspots to provide data to authorities to implement lockdown measures earlier. Although both these models take a more abstract approach in moving away from individuals and more to mobility/populations trends this doesn’t indicate the analysis provides any less value.

Data Acquisition Techniques

Linked closely to study location, data acquisition is a field for which numerous techniques are used for a variety of systems, which can be applied to a range of situations.

Yu Zheng et al. [14] looked to infer GPS movements from individuals GPS logs. Due to GPS's highly accurate nature, the team looked to not only predict movements but to infer motion modes (transportation use etc) with a dataset of 65 people over a period of 10 months. The survey was constrained to this population size, highlighting the main issue of using GPS techniques related to privacy constraints (consent and legal implications). However if the data can be obtained the accuracy and consistent temporal updates that it provides at the scale it is used within is unrivalled using current modern techniques. The study found, that to infer movement types through features such as heading, velocity and stop change rate an accuracy of 72% was achieved.

Another technique used with the larger location-based studies is the use of cellular data which is only generated when a phone is used for a voice call or text-message [20]. The article explores the increased use of cellular datasets from its inherent ability to collate millions of records through telecom provider logs. They also result in very little privacy or ethical issues during the collection process. The paper also discusses the deep-rooted and unalterable flaws of temporally sparse, low-accuracy data. Yves-Alexandre de Montjoye et al. [16] conducted research with 1.5 million individuals (1/4 of the population) over a fifteen-month period using records from telecom provider Orange within the Cote de Ivory. Although this paper outlined the privacy limitation of such data, it gave an insight into how the data was used and implementation strategies for data lacking accuracy. Population flow could be tracked, and workday-weekend cycles could be analysed, indicating that although pinpoint accuracy cannot be obtained, a more abstracted view of population movements can be formed.

To adapt to the need for smaller-scale human mobility, GPS and Cellular data are considered ineffective for lower scale individual-based studies, as in urban areas where an individual will spend 80-90 per cent of their time indoors, crippling their effectiveness [22]. In these situations, WiFi-based solutions are employed although implementation varies considerably. G Biczok et al. [22] used an IPS (Indoor Positioning System) called Maze map which employed WiFi trilateration, where signal strengths from 3 or more different access points are measured (to an accuracy of 5-10 metres) and GPS was used when necessary, to fill gaps. Data was presented through the Maze map App that was used on multiple university campuses and produced a depersonalised (privacy preserving) positioning log of each user that conformed to EU privacy law. In contrast, Piotr Sapiezynsk et al. [29] used a data set of only 63 people but the position sample was taken every 16 seconds allowing for extremely accurate spatiotemporal data. The aim of this study was to find a percentage of positional data that can be obtained through WiFi access points within an urban setting, the research concluded that access points record 80% of mobility. This also demonstrated the functionality of WiFi access points and with the addition of GPS, to provide an effective solution.

Radio Frequent Identification Devices (RFID) has been used more recently to present a new technique in which accurate indoor datasets can be produced whilst providing person to person interaction data. RFID is a technique that uses active mutual device proximity to obtain highly accurate spatiotemporal data. Andrea Cangialosi and Joseph E. Monaly [21] tried to implement an RFID system into a more enclosed setting of a hospital to benefit patient care and hospital operations rather than a more traditional human mobility solution. They implemented RFID frameworks, but their goal was more to illuminate the multiple possibilities of this technique and how it can be beneficial for an organisation to employ. These benefits came as operational process improvements (staff assignment, asset management, patient movement and hospital billing) but this came with a large set of operational and implementation issues. It was found that even though, no other system could provide such in-depth analysis no other system had such a large start-up cost and the need for operational upkeep. The key issue with RFID implementations was the lack in pre-existing systems from which data could be produced, this caused large initial installation costs pushing away potential future research. I agree with this paper’s analysis of RFID and the benefits that it can provide but other uses need to be developed if the RFID solution is to be affordable.

Privacy Considerations for Human Mobility

Although not a key part of research when considering model creation and analysis, privacy considerations and the related ethics are always needed to ensure data can be as anonymous as possible.

A foundation of this research within human mobility was conducted by Yves-Alexandre de Montjoye et al. [16] the goal of this study was to find the number of spatiotemporal points needed to uniquely identify an individual using a “unicity test” model. The motivation behind such a study was to highlight the privacy issues that anonymous datasets paired with external publicly available data (voting records, home address, geosocial media interactions etc.) can cause. The study found that to uniquely identify 95% of individuals only 4 spatiotemporal points were needed, this implied that when designing human mobility frameworks, privacy considerations need to be explored thoroughly. This presented a real challenge to being able to undertake accurate mobility analysis whilst protecting individual identities.

Roberto Pellungrini et al. [30] also investigated the privacy risks of re-identification of individuals within their framework design but attempted to solve this by mitigating the current shortcoming of existing mobility frameworks [31]; computation complexity and recomputing privacy risk for dataset changes. The idea was to train a classifier to not only capture human mobility relationships but to also measure the amount of privacy provided by it, this was done through a thorough risk assessment and a process known as

Privacy Risk Computation to simulate attacks (based on the assessment) on the dataset to compute a risk value for a model.

Daniel Soper [32] looked to analyse the ethical implications of human mobility research and the trade-off between mobility technologies and an individual’s privacy. As explored above, the paper found that although privacy can be improved upon in some sense, human mobility data will always be susceptible to privacy exploitation. Considering this, Soper’s analysis examined the social benefits that research could provide whilst presenting the potential exploitations of which governments and corporations were felt to be the most likely to exploit the systems created. The work concluded that there is a fine line when protecting an individual’s identity and data whilst not impeding the overriding social benefits of research. To resolve this research indicated the critical role that Wireless Service Providers (WSPs) would need to play with them needing self-regulate data with a universal set of base principles whilst operating within legislation. To link back, it is very important for research in this field to also use such principles during dataset collection and to stay clear of unethical practices.

Synthesis/Research Question Creation

The key in reflecting on the literature review was to establish a topic gap for this research. Although quite abstracted, the review of the current research literature has provided an insight into the current methods of analysis within the field as well as indicating current research downfalls and study gaps, sufficient to be able to develop the research requirements/questions for this thesis and the associated project goals. As a research area, human mobility modelling has focussed on improving the understanding of an individual’s movements but research has failed to provide sufficient analysis to classify groups of individuals based on their movements and habits. This has meant that it has not been possible to either predict their current task or to identify the movement behaviours of compatible people within a system (e.g. store workers, school teachers, school students etc.), based on their total behaviour and movement trends. To move this forward, a research question has been focussed on how to categorise agents and identify known groups.

Another shortfall within the field has been the lack of a universal framework to implement WiFi and RFID techniques. Such frameworks have been wildly created for GPS and Cellular data-based studies some of which have been discussed above. Recognising the paucity of RFID datasets and the associated costs and issues of implementation, RFID has been discounted which has led to the second research target to examine how WiFi traces be adapted to aid classification and categorisation of groups.

It was clear from the analysis of the papers, that provided an abstracted look at the scenarios they were analysing (offering functionality to be applicable to multiple scenarios) had larger uptake in practical applications and further implementation. Although framework solutions were explored within the literature study, all provided convoluted answers to the problems they were facing. As this is a key aspect of this type of approach, the third clear research target was focussed on the feasibility of a framework solution that will try to provide a flexible implementation which lead to a specific research target on adapting a framework for other Systems/Locations.

Further, other popular research has produced outputs that are very interpretable to the reader, so that they can easily be used in further research. Considering the output of the solution produced from the other research questions, analysis will be conducted to produce usable outputs that not only inform users but provide analysis to easily implement further study on top of this research. To assess this implementation, supports the fourth research target to assess what the analytics produced from a model can be used for.

To conclude, the literature review has proven invaluable in supporting the investigation into the research area and has given insight into the research questions that have been chosen to evaluate the solution later in development. It was also key to finding the topic gap "Group classification within a mobility scenario". The Research Questions that will provide the foundation of analysis for this research are in [Table 1].

RQ1	How can we classify agents and identify known groups?
RQ2	How can WiFi traces be adapted to aid classifications?
RQ3	How can the system be adapted for other Systems/Location?
RQ4	What can the analytics produced from a model be used for?

Table 1 Summary of Research Questions

Project Breakdown

Data – a validated open source dataset will be utilised and form the basis of the analysis. This choice will be validated against a selection scheme.

Method/Development – this section will explore processes used in other research to then find a solution that can be implemented to the classification problem. Will provide a software-based solution that will use industry standards, methods and protocols.

Results – this section will outline the results produced by implementing the method as well as drawing conclusions drawn from the data provided. This section will support the in-depth analysis on how the solution can be used on a more practical level from real-world implementation.

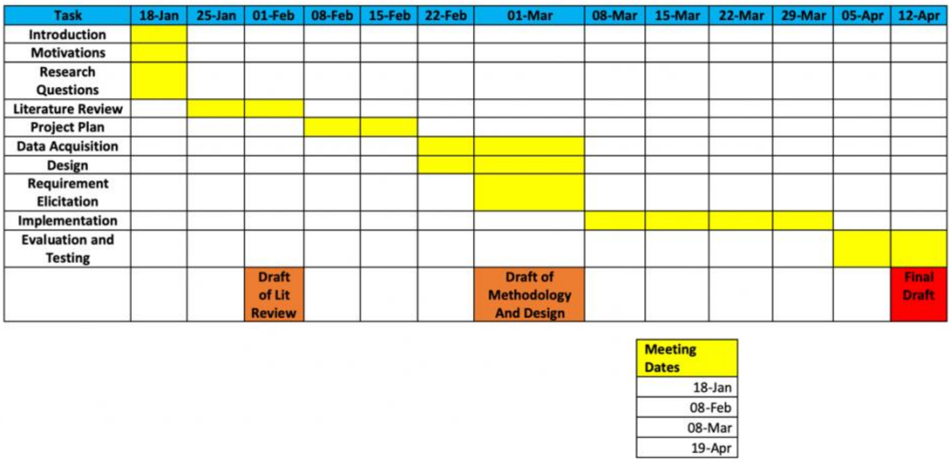
Conclusions – this section will provide develop the analysis into a number of key outcomes, specifically the 4 Research Questions and identify areas for improvement.

Future Work – this section outlines how this work can be built on, and prompt areas for further investigation.

Project Plan

The Gantt chart below shows the details and timeframe of the project plan.

Gantt Chart -



Task	18-Jan	25-Jan	01-Feb	08-Feb	15-Feb	22-Feb	01-Mar	08-Mar	15-Mar	22-Mar	29-Mar	05-Apr	12-Apr
Introduction	🟨												
Motivations	🟨												
Research Questions	🟨												
Literature Review		🟨	🟨										
Project Plan				🟨	🟨								
Data Acquisition						🟨	🟨						
Design						🟨	🟨						
Requirement Elicitation							🟨						
Implementation								🟨	🟨	🟨	🟨		
Evaluation and Testing												🟨	🟨
			Draft of Lit Review				Draft of Methodology And Design						Final Draft

Meeting Dates
18-Jan
08-Feb
08-Mar
19-Apr

Methodology

Functional Requirements Elicitation

The requirements for this study were created to ensure that development did not fall short of the Research Questions (RQ) formulated through the literature review. Each RQ will enable requirements to be developed to formulate the research task, to achieve the desired output.

Research Questions Overview

RQ1 - How can we categorise agents and identify known groups?

> The central objective for this paper is to create a model that can categorise groups of people within a system based on human mobility data. Such modelling could be either supervised or unsupervised depending on the chosen dataset and its limitations. Due to limitations within the dataset, features summarising the data may need to be produced to help ease the categorisation problem. The model will also need to be optimised and data processed to reach the solution. This was a significant task to ensure that the baseline analysis and development of the subsequent RQs could be achieved.

RQ2 - How can WiFi traces on be adapted to aid classification?

The goal for this question was to provide insight on how to manipulate WiFi connection data to allow for easier categorisation of individuals that are connecting to it. This could involve the process of applying a range of techniques to achieve the best functionality from the data and facilitate the subsequent analysis. Further features will also be assessed to improve the analysis of individuals within the system because singular WiFi connection will provide very little on its own but in larger frequencies can provide a deeper understanding of the scenarios.

RQ3 - How can the system be adapted for other Systems/Location?

The goal here is to create a method that is not only implementable in the location the dataset was used but can also be transferred to other systems in which other models can be formulated in similar ways changing only what is necessary to accommodate the trends in the new location. Projects explored in the literature review opted for an abstracted open function to provide a generalized solution, but other aspects should be explored.

RQ4 - What can the analytics produced from a model be used for?

Following on from the enabling RQs, this section focuses on how the approach and analysis framework can be utilised for educational and commercial uses. It will also assess approaches for incorporating privacy bounds and real-world implementation problems. This will be a post system build analysis task that will look to justify this software and look for the benefits of using it to aid people to make more informed choices based on mobility data.

The RQs were broken down the problem into 4 areas to shape the development of the project. A breakdown of requirements based on the MOSCOW requirement framework is at Table 2.

ID	**Description**	**Priority**
FR_MODEL_IDENTIFY	Model must be able to identify and classify the known groups within a system.	Must
FR_MODEL_DATA	System should be able to process mobility data passed to it to produce features that can be used to both train and test the model	Must
FR_MODEL_TYPE	Model should be able to categorise groups in an unsupervised way	Must
FR_MODEL_TESTING	Model needs to have the ability to be tested and for it to be ranked with others.	Should

FR_MODEL_ HYPERPARAMETERS	User will be allowed to make changes to hyperparameters.	Should
FR_DATASET	System should use a dataset that uses WIFI connection logs and has identifiable groups within the collection area	Could
FR_MODEL_ ADAPTABILITY	Model should be adaptable to allow for other scenarios/location of the same data structure to be analysed e.g. supermarkets etc	Must
FR_ANALYSIS	Insights to be provided on how such analysis can be used in a commercial/educational setting	Must
FR_MODEL_ EXEMPLAR	Model should produce an exemplar agent that would be located at the exact centre of each cluster	Should

Table 2 Functional Requirements

Data Search/Analysis – kth/Campus (Dataset)

When selecting the data for this research the creation of a research defined dataset for a WiFi traced human system would not have been possible due to the complexity involved alongside the ongoing coronavirus pandemic. Therefore, an alternative was found, using the CRAWDAD dataset library [33]. The CRAWDAD community platform acts as a hub for all data linking towards research within the human mobility field among other closely linked research areas. CRAWDAD is a Dartmouth University [34] sponsored program that enables the sharing of datasets within this research community. Datasets from CRAWDAD also come with some intrinsic benefits, such as pre-prepared data cleaning tasks (null values & noise removal) before datasets are submitted to the community. CRAWDAD data also comes with an academic free use policy with no constraints on its use, all information of the libraries free use policy can be found at CRAWDAD’s data license agreement [35]. CRAWDAD does have a policy restricting the use of de-anonymising data through re-identifying processes, which could limit some research but will not affect this paper.

The following key principles were used when identifying the optimum dataset within CRAWDAD, these were elicited in compliance with the RQs:

- Anonymity – data should adhere to EU privacy standards and must be pre-anonymised, to enable ethical analysis and practises.
- Scale – the dataset size was taken into consideration with smaller and exceedingly large datasets being ignored due to complexity issues and challenges with computational complexity.
- Acquisition Technique – the acquisition technique used to acquire the data indicates the scale of the dataset as well as data quality.

This project needed a dataset that could ideally represent mobility trends over a group/groups of people whilst providing enough data to enable more interpersonal trends. As stated within the research question RQ2 a WiFi acquired dataset will be used as it provided the optimum balance between accuracy and obtainability.

- Scope – datasets needs to have several identifiable groups that were recorded within the dataset whether they were labelled or not.

Using the above strategy, the kth/campus dataset was chosen, this dataset was acquired by Ljubica Pajevic, Viktoria Fodor and Gunar Karlsson from KTH University in Stockholm. The dataset covered the 5 campuses of KTH University within metropolitan Stockholm seen in Figure 1, using WIFI access point connection logs to monitor mobility. The coverage of the Wireless access points meant that not only were indoor associations made, but also covered outdoor areas with good accuracy. Furthermore, this dataset's scope and chosen acquisition technique suited the categorisation portion of this study by providing enough data that can be adapted to implement a machine learning model around it. This also fulfilled the FR_DATASET requirement thus partly answering RQ1 & 2.

The dataset was originally used for the analysis of device/user connections and trends of usage within a dense Access Point (AP) network. The dataset consists of two tracesets (Eduroam-traceset1 and wifimapping-traceset2) for which only the Eduroam-traceset1 will be used. Eduroam_traceset1 acted a log of all connections between devices and the WiFi access points on the campus, the associations were recorded with a timestamp for each individual connection between a device and an AP, with the data set providing a timestamp, hashed client ID (users device) and AP identifier (e.g. Bldg1AP1). Within the dataset AP's we are given x,y coordinates to allow for the mapping of individual AP's with campus locations but due the uses of such data disconnecting with the research questions scope, it was ignored but could aid further research. Each AP was identified through an ID providing information on building name, floor, room and corridor. The data was acquired over the period of 16 months for which the number of AP's varied from 930 – 984, over this time period, the average number of unique users daily was 1400 from a pool of around 18000 users who could connect to the university exclusive Eduroam AP's. Data acquisition was collected through a Radius Logs acquisition technique outlined by Ljubica et al [36]. Radius provided a simple but effective method to authenticate users in which triggering a new association with an AP at specific points, these re-association protocols kept data accurate and up to date.

When assessing the applicability of the data to the problem, it offers two identifiable groups - university students and university staff. But as can be seen from the dataset no categorisation was recorded and as such there became an opportunity in which an *unsupervised model* could be trained to identify these groups within the system. This model would need to train this classification through the spatiotemporal points provided as well as features that could be pulled from this data for each user. The first overall feeling from the dataset was that these two known groups could be differentiated due to different spatial footprints as well as differing behaviours such as how sporadic each group is on campus and the consistency of connections.

Hypothesis

When solving the problem of agent classification within the above dataset an initial hypothesis is needed to provide a preliminary belief on how each group will behave. Having defined two agent classes that are present on the campus, staff and students, I consider that although these groups will exhibit deep-rooted regularity with their mobility as found in [5] there will be some key differences that will allow for differentiation between the two.

Firstly, when considering staff movements, I am confident that they will follow regular movement patterns that show little random or sporadic movement, in the most part due to the fixed location and nature of the roles often centred around single locations in faculties, similar in part to patterns exhibited by office workers [37]. Students on the other hand will exhibit more random, less predictable and sporadic movement patterns, covering more of the campus due to their academic schedules which will revolve around going to lectures/practical-demonstrations which take place in varying locations. There is also an added social mobility that students will be likely to exhibit, causing random mobility patterns to be seen. I believe that they will represent a more pedestrian movement pattern, also seen in [37]. I have a concern that although these behaviours will be present within the data, due to the continuous nature of this data no visual clusters will form and the data will have a large single footprint.

This hypothesis can be evaluated using the techniques described in the ‘Testing Techniques’ section. In short scenarios will be found that have different groups present to provide ground truth data and the clustering schemes will be evaluated against them in both shape, time and density.

Data Implementation/Pre-processing

As outlined in RQ2, WiFi logs need further processing to produce and extract the correct information, as clustering in its raw pre-processed state provides little insight to group behaviours. The only realistic and feasible solution to this problem is to use data features that summarise an individual’s connections to the APs on the campus. The creation of these features will help to answer RQ2 by providing a set of features that can be used on all WiFi logs of the same structure and in doing so extract mobility information.

Before features could be extracted from the dataset a timescale for the attributes needed to be provided for which 2 solutions were apparent, either features could provide information on daily movements or an overview of weekly movement patterns of an agent. Monthly features were also considered but discarded due to the large computation time that would be required. These two choices provided different options for model design and focus for the project. The final decision on using weekly analysis was based on reducing the computational cost of the system whilst being able to summarise enough data to nullify data anomalies and create a fuller and more accurate assessment of each individual's mobility.

When creating the summary features, only 3 columns were able to have features drawn from them: timestamp, apID and XY coordinates/floor (dataset attributes). All others provided no additional information to basic record identifications which does not add to the group or individual classification and analysis. From the 3 columns, 7 features were produced to embody the hypothesis so that sporadic users could be differentiated from those that had more routine movement patterns:

- Unique Associations – will identify agents who tend to spend time around the same APs regularly, identifying those with a large and small footprint around campus.
- Present On Weekend – different groups such as students and staff wont be around on weekends (for the majority) and so will provide insight on the individuals that are and the groups they are a part of.
- Hours On Campus – gives an overall view of time spent on campus.
- Total Connections – will give an idea of how long an individual has been on campus, and the number of connections they connected to
- Unique Buildings – calculated using the apID and will look to target the number of areas the agents has visited within the system.
- Sporadic Metric - Total Connections/UniqueAssociations - will pinpoint agents that centre around certain locations and will increase with how sporadic an agent is.
- Days On Campus – should be able to provide supplementary information to help clustering

Data Processing

Data processing is used to improve the effectiveness of machine learning modelling techniques. The implementation of these different methods were analysed to further answer RQ2. The majority of pre-processing techniques analysed in the field involve the three methods explored below, these offer increased functionality by building on the previous techniques to provide further insight into the data. These will also cause the machine learning models to behave differently, which will be evaluated using the models defined in the unsupervised modelling section and evaluated using the testing scheme defined in the evaluation techniques section.

Correlation Matrix Analysis

Before any modelling takes place and all the data has been provided, a feature correlation matrix will be created to support initial analysis and look for correlation between features. Based on this correlation, features will be removed (if two correlate with each other) or more will be added if there is little variability within the dataset.

Principle Component Analysis (PCA)

Before PCA the model created from the data will use all the data no matter how useful it is within the classification problem. To solve this PCA, will be used to not only centre the data but if needed will provide dimension reduction which could lead to an improved interpretation of the data and produce better classification. Furthermore, it will allow for easier visualisation of the data such that it can be plotted and data trends can be analysed more easily. This will clear up any messy clustering classifications if they occur.

Normalise the Data

Now that these features have been created a normalisation function will be needed to allow for all metrics to adopt the same scale and to improve classification later. The normalisation technique used will be that of a MaxMin scaling as seen in Figure 2. The improvement that could be seen through the implementation of a normalisation scheme is that of a changing variance within the data. This change will mean that different features will inform the clustering more and such new trends might be seen as well as an improved classification.

Unsupervised Modelling

When looking to train a model for an unlabelled dataset to answer RQ1, three overarching unsupervised clustering techniques were explored: Centroids-based, Distribution-based, and Density-based. Others techniques were assessed including fuzzy and hierarchical clustering but were not utilised because of research constraints and poor functionality for this research. The clustering techniques were analysed for their strengths and weaknesses including their theoretical and practical characteristics, how easy they are to implement and the accuracy that each technique provides when subjected to dissimilar data shape. The goal when selecting models to use for the classification was to implement the varying clustering techniques described below and assessing their performance against RQ1.

Centroid Clustering - K-Means

K-Means is a standard clustering implementation that solves the classification problem with the goal of locating the centre point of each group within the dataset. It requires the user to select how many groups the algorithm should classify, this is generally considered to be disadvantageous but this was not the case for this application because these algorithms are considered very quick in contrast to others. Furthermore, out of all clustering algorithms, this is by far the most used and much is known about how it behaves when subjected to different dataset variations. K-Means works by repeatedly calculating the mean value of a selection of points within a range, with each iteration updating the mean and classifying new point until a plateau is reached. In effect this technique iterates to a new datum point.

K-Means employs distance-based similarity metrics (Euclidean and Manhattan) which look to find similarity based on location between pairs of data. Other options are available (Pearson's Correlation, Jaccard Similarity and Cosine Similarity) but these were not explored within this methodology, although research has shown that they can lead to boost in clustering performance [38].

Distribution Clustering (Gaussian) - Expectation Maximisation

A more flexible option for unsupervised clustering, is forming clusters around a probability distribution. The most popular method for this clustering implementation is called Expectation Maximisation. This method like K-Means needs the number of clusters to be provided, and for each cluster a random location, distribution and mean is used to initialise the process. From this, each data point has a probability of being assigned to a cluster (the closer the point is to the gaussian centre the higher the probability). This technique comes with a higher implementation and computational complexity cost but does provide more flexibility than that of K-Means as more elliptical clusters can be formed and such matching data with more accuracy. As the shape of the data could be in the form of a large cluster, this will allow the scheme to produce results and find trends where visibly non are shown. There is also a hope this will lead to an alternate classification that will provide clustering options to increase classification accuracy.

Density Clustering – DBSCAN

DBSCAN is a density-based clustering system, which is similar to mean-shift (a version of K-Means) but measures the density of points as opposed to distance. The process works by checking data points within a window and if a pre-defined/certain amount of data points are found clustering begins.

Clustering involves adding data points to a cluster that are within a window and then iteratively shifting that window onto those points. This method thrives on clear non-circular clusters that are visible and as such, if the hypothesis is correct about the data’s shape this will struggle to provide any meaningful insight but is included so that if the hypothesis is wrong, there is a fall-back option.

Software Solution

To implement the above clustering methodology, a software solution needs to be created. The plan for the solution is for an input to be provided in which the system will automatically process the data to a standard where it can subsequently be used for cluster analysis. The system will then be used to learn from the data and in turn separate the groups within the scenario. The algorithms will be tested to see how accurate they were and then give the best clustering technique for the scenario. At Figure 3 is a flow diagram describing the process that will be used to process the data and then create models from it.

To implement this software solution I chose to utilise Python because it is a language that I have used extensively, it is well-suited to the large amount of processing that needs to be done with the dataset and there are a number of relevant libraries (scikitLearn, Pandas, SciPy ands NumPy) to aid the implementation of the clustering algorithms.

Evaluation Techniques

To test the outcome of the models created through the system, two techniques could be utilised, Metric based and Ground Truth Evaluation. The first of these techniques uses evaluation metrics (Rand Index, Folkes Mallows Score and Silhouette Score) to measure the quality of the clusters created by the algorithm. Metrics were considered ineffective in this scenario due to the lack of evidence to suggest clear visual clusters for which these techniques thrive. Due to this ground truth datasets will be used to inform the evaluation and see not only if the clustering shapes are correct but also the ratio of agents classified is also good.

Ground Truth Evaluation

Before justifying further, a definition is needed for what a ground truth set is: it is information that provides evidence of something that is fundamentally true within a situation. In the case of this research ground truth data can be considered data from a timeframe within the collection period that represents a specific group that is known to be on campus.

To formulate these ground truths, timeframes were found within the dataset, where the location of entire groups present on the KTH campus were known.

Due to the constant presence of individuals on a university campus, there are very few timeframes that can single out certain groups. To solve this problem, 3 datasets were defined to provide enough information to either provide an empirical ground truth or provide enough information that ground truths could be inferred.

The first ground truth was formulated from data within the Christmas and New Year week starting 24/12/14 (dd/mm/yy), this provided a window of time in which only non-moving agents were present on campus. This allowed an empirical ground truth to be constructed in this time as it is not blurred by the other classification groups. This was confirmed through the KTH website [39] that stated that both students and staff are prohibited to enter campus around this time without authorisation due to security reasons. This use of constrained data is a valid ground truth data point for subsequent analysis.

A similar method was used to create a ground truth including non-moving agents and staff connections. Using the university student calendar, a time frame was selected where staff were active on campus and student were absent. The data selected was for a week commencing 09/06/14 immediately prior to the start of the university year with staff on campus. Due to this data containing individuals from both the non-moving agent and staff groups further refinement was required.

This refinement involved examining the footprint of the majority of the ground truth data for the non-moving agents, then finding the difference between that and the footprint produced for the joint ground truth of non-moving agents and staff connections. This in theory should indicate the shape of the staff data within the dataset and provide the system with a ground truth to evaluate clusters of the staff class.

A comparable technique was then used to create the ground truth for the student class. Due to the mobility behaviour of students on a university campus, there were no time windows where students were active on campus and staff were absent. To this end, a time was selected to encompass all groups present on campus. The time frame chosen was a week at the start of the university year a few weeks after student induction beginning 08/09/14 to maximise populations for each group. The refinement technique used above was then applied but instead the difference was calculated between the ground truth data set representing both the non-moving agents and staff connections and the data encompassing all groups. In theory this should indicate the shape of the student data and provide a ground truth so evaluation can be conducted.

Ground Truth Sets

24/12/14 – Christmas Holidays – Non-Moving Agents
Total UserIDs - 3566

09/06/14 – Before Start of Term – Staff & Non-Moving Agents
Total UserIDs - 11097

08/09/14 – Start of Term – All groups present
Total UserIDs - 23833

Implementation Challenges

Pre-processing

During the pre-processing stage of implementation, the Pandas library proved invaluable when formatting the data. To initially format the dataset a test dataset was used that consisted of 12 days of data at the beginning of the experiment. This data was used to test how the data was formatted and to provide base functionality to then build-up to the larger datasets. When looking at the data provided from the KTH research, it was split into monthly intervals from January 2014 to April 2015. To assess the data more efficiently, weekly datasets were formatted, this provided large amounts of functionality within my code allowing for easy retrieval of AP associations from a given period. Figure 4 is summary of the dataset for the week starting 9th June 2014.

Data Scale

When analysing the data set, the large scale of the data that was being worked with became very apparent. Due to an oversight within the methodology, issues regarding the scale of the data were not explored and such special considerations were needed to explain how this data would be tackled. The good news, when referring to the scale of the data, is that the number of users that are recorded within the system is a lot less than that of the total connections within a week. Furthermore, the number of users also changes based on the week that is being analysed and how busy the campus is depending on the time within the university year. In the above week, 11059 unique users were present with some weeks in the height of the academic year having nearer 25000 unique users. Even though the reduction of data meant that the problem of data scale wasn't so vast as once thought, features creation was still required.

Feature Creation

The process to create the features discussed within the methodology was undertook in Python and the Pandas data formatting library was used to help provide data manipulation methods. During this process, all of the features were created using test data provided within the dataset. During testing, times were recorded to see how each feature would perform and how well they utilised the computer hardware. All features performed well computationally with each feature taking upto half a second to produce.

When computing the HoursOnCampus feature, the volume of data and associated computation time meant that some tasks took between 20 and 30 seconds for each user. With this being such a large portion of each user's features computation time, analysis began to look at ways of reducing this. After much deliberation and implementation of time reduction programming techniques, a decision was made to remove the feature from the computation as times were still around 15 seconds per user and such would require up to 5/6 days of computation with the addition of HoursOnCampus feature. Whereas when computing the other 6 features for one of the larger datasets the overall computation time took just over 26 hours.

Due to the large feature creation times, only one testing feature set was created with its goal to see the consistency between the data and make sure no odd classifications occurred with the final clustering cluster scheme. This data was defined as follows:

> 03/05/15 – Middle of Term – All groups represented. Total UserIDs – 24567

Results

The goal for assessing these results was to evaluate the classifications produced to answer RQ1 and the chosen clustering algorithm. Depending on how well RQ1 is met, this will affect RQ2, and proving that the techniques used have either aided or detracted from the classification. Within the data processing segment of the methodology, processing methods were applied in the order they are defined and testing was conducted to measure their effectiveness in order to answer RQ2. The 08/09/14 feature dataset was used for the clustering as well as acting as the student ground truth data set.

To initiate analysis a Correlation Matrix was implemented. This provided little insight, so much so that it was considered ineffective when evaluating data. This was due to the small number of features within this data set. Nonetheless, some insight could obtained, highlighting the high correlation (0.87) between the uniqueAssociations and uniqueBuildings features. This insight suggests that either could be removed with very little impact on clustering, an example of which is at Figure 5.

Initial Clustering

To assess these results, plots needed to be created for both the clustering classifications and the ground truths. The features uniqueAssociations (x) and totalConnections (y) were selected due to their high variance suggesting higher chances of clearer clusters forming, each features variance can be seen within Figure 6. The clustering methods were all trained using all available features.

To evaluate this initial clustering the ground truths were plotted as shown at Figure 7, Figure 8, Figure 9 and were used as defined within the methodology. From these plots, analysis of behaviours between the groups could be recognised. In Figure 7, non-moving agents can be seen to have a lower number of both totalConnections and uniqueAssociations than the other classes. From the hypothesis non-moving agents should only have one unique association but the data shows individuals connection to up to 40 unique APs within the time frame. This can be explained due to the dense AP network within the university with non-moving agents connecting to multiple AP within its vicinity. Furthermore, the larger numbers seen, can also be justified through anomalies within the data such as rare staff and student visits during that period.

The other grounds truths also showed initial support for the hypothesis with the students having larger uniqueAssociations thus implying a larger sporadic footprint. An interesting side observation was the indifference between the maximum values seen in both the student and staff ground truth thus inciting a need for a change to plot features to help aid classification.

K-Means

The clustering classification K-Means produced can be seen at Figure 10, when analysing the clusters, trends could be identified. The clustering scheme seemed to not recognise the relationship between staff and students. But it was able to distinguish between the non-moving agents and the other classes with both data point density and footprint but due to the linear classification I don't believe that this is a correlation, I judge it to be more of a coincidence and an anomaly of the classification technique.

Gaussian Mixture

The Gaussian Mixture worked in less linear way when compared with that of the K-Means with it correlating data from other features (causing classes to overlap, as data was clustered on a different axis). No great insight could be drawn from this clustering scheme and the classifications we very inaccurate when evaluated against the ground truth. This is shown in Figure 11.

DBSCAN

Due to the form of the data that was given to the algorithm (i.e no discernable shapes), the classification only recorded one class within the data. This was by far the worst of the clustering schemes with clustering all the data into one class. As stated within the methodology this algorithm relies heavily on visually distinct clusters which this data is struggling to produce and such until this happens this will achieve poorly correlated results. This is shown in Figure 12.

For an initial introduction, the insights provided by the ground truths were invaluable in beginning to answer the RQs. This was no more apparent than showing that the features produced to summarise the WiFi connection logs could aid classification by producing data that reflected the behaviours defined within the hypothesis.

Clustering After PCA Introduction

PCA was introduced to improve upon the shortfalls of the previous development iteration. Inadequacies included the lack of accurate clustering due to the data's shape and the use of certain features not differentiating the data to allow the ground truth to reflect the hypothesis.

The introduction of PCA should provide dimension reduction to remove the less influential features from the classification equation, easing the problem whilst not losing important traits within the data. Furthermore, the implementation of PCA will remove the behaviour shown in Figure 11 with clusters that seem to overlap, consequently removing noise and increasing the predictive power.

Once again analysis was conducted on the ground truth data seen at Figure 13, Figure 14 and Figure 15. When comparing it to the previous iteration the ground truths were much less interpretable due to the data centring that was implemented through PCA. Nonetheless, the behaviour and footprint of each class could be identified, but with the ground truth clusters being messier than in the previous iteration.

Clustering Evaluation

Due to the nature of the ground truths and the fact that they produced very messy/unbounded footprints, it was difficult to evaluate the clustering schemes. The K-Means classification as shown in Figure 16 produced a clustering scheme that in no way reflected the footprint shown by the ground truths. As stated in the previous iteration the linear clusters formed from K-Means started to show weakness when classifying more complex scenarios. This also confirmed the suspicion that the initial K-Means classification was not representative of the success and validity of this technique.

The Gaussian mixture was then tested with the classification shown at Figure 17. When considering this iteration as a whole, this classification reflected the ground truths the best. Although not perfect when classifying the non-moving agents 90% of the density recorded within the ground truth was classified correctly. When looking at how the scheme clustered the student and staff classes, errors started to occur such as a large over-classification of staff and the under-classification of students.

DBSCAN was finally evaluated and just like in the previous iteration this (shown at Figure 18) produced very little and clustered the majority of the data as one class. Due to the null results that this algorithm was achieving over multiple iterations, a decision was made to remove this clustering scheme from the rotation and implement a new algorithm much more suited towards the data shape. BIRCH was chosen for its ability to provide alternate clustering in similar scenarios/data shapes where no discernible clusters can be found. This was the then analysed on this iteration with no apparent results (shown at Figure 19) but previous research indicates future classifications being more fruitful.

Introduction of Normalisation

One limiting factor found within the previous iterations was that the shape of the data not fully reflecting the hypothesis, this was caused due to the features used to plot both the ground truth and clustering results. The axes were previously selected via variance calculations made from the initial data set. To build upon this, normalisation was implemented to provide a mix up within the data and to track the changes that occurred.

The first noticeable change was the variance of features within the data, shown at Figure 20 where there is clearly a significant shift of values. This then directly affected the PCA dimension reduction with the principal components changing and providing new shapes within the data. The principle components created from this implementation were the features of DaysOnCampus and presentOnWeekend.

When assessing the data, the graph produced can be seen in Figure 21. This in theory was ment to provide large variantion to aid cluster creation. Although this did happen, and clear clusters had formed, the data in no way represented the hypothesis and the mobility of users within the system. This is due to the original purpose of the features chosen, being that of supllimentary features that would help the classification but not provide key insights on behaviour of individuals.

To amend this the presentOnWeekend feature was removed from the data set with this producing the data plot at Figure 22. Due to the nature of the discrete features that were still informing the clustering, the data still struggled to reflect the hypothesis and differentiate between each ground truth.

Due to this continued problem the daysOnCampus feature was removed with the result shown at Figure 23. PCA dimension reduction then left uniqueBuildings (x) and uniqueConnections (y) as the features informing the classification. This made an immediate impact with the ground truths that can be seen at Figure 23, Figure 24, Figure 25 and accurate portrayed the mobility footprint of each class.

When analysing these ground truths, comparisons could to be made between previous ground truths and this set. Within the first iteration the ground truths provide a great initial insight into behaviours but fell short through only providing one feature that could differentiate between classes. These ground truths support the hypothesis with non-moving agents at Figure 25 seen to carry low mobility within the system with steady number of uniqueConnections and the majority of the data point density centred around the bottom left hand corner. Once the staff and student ground truth were inferred they almost exactly represented the hypothesis defined within the methodology with gradually increasing mobility as data points moved away from the bottom left corner.

Clustering Evaluation

The K-Means clustering as shown at Figure 26 produced nothing of importance, with clustering forming with very little correlation to the ground truth. BIRCH clustering at Figure 27 was also implemented but with the same problem, that it was unable to find the relationships with the data.

The Gaussian mixture classification can be seen at Figure 28. This clustering recognised the mobility relationship between each of the groups. This was cross-examined with the ground truths and although the footprints of the classification didn't mimic the footprint of the ground truths there was a close resemblance. After further examination of the density of data points within each of the ground truths and then re-examine the clustered data, using basic problem analysis the 'misclassifications' could be explained. When comparing the footprints between the ground truths and the clustered data it would be easy to say that although they mimic each other they aren't very accurate. As briefly explored above, insights can be made when looking at the density of data point populations to explain that the majority of classifications are correct

Using the density of data points within each plot helped to confirm that the majority of the points were where the classification had identified them within the clustering method (shown as dark blue) in Figure 28. Other points not within the small area in the bottom left-hand side were either anomalies or maintenance staff moving around campus causing our ground truth set not to be 100% accurate. Within the non-moving agent ground truths set seen in Figure 25, it was a little different with two classes being present within the plot but although it took up a large footprint the majority of the data was contained within the two areas which were represented in the classification. Once again, the excess data points were anomalies within the dataset that could be described as students who were on campus not during term or staff with increased sporadic movement.

Final Clustering Scheme

The final clustering solution [Figure 28] that was settled on for this classification was that of a Gaussian mixture algorithm, that utilised the paper defined uniqueConnections and uniqueBuildings features that had undergone PCA centring and normalisation.

When looking at the final clustered solution the classification was that of the dark blue representing that of the Non-moving agents (printer, staff laptops, etc) unwanted associations within the system, white then represented the data points of staff within the campus and then the students will be represented through the light blue. To then test that this classification worked over multiple datasets, clustering was the ran on features from the week starting 03/05/15 which resulted in the following classification Figure 29. As can be seen, by the plot this clustering almost identically replicated the one from the 08/09/14 dataset. This not only proved that this clustering algorithm was versatile, but it also showed that the original one was a one-off prediction.

As can be expected from such an unsupervised approach to classification this will never provide a hundred per cent effectiveness for a dataset and so some assumptions had to be made on selecting the final clustering solution. Even with the evidence from the ground truth data set, it would be naïve to think that staff and student's footprints within the clustering are not as cookie-cutter as the clustering might show. Due to this, misclassifications will not be common but must be expected as anomalies and incorrect clustering's will occur but with the problems that an algorithm like this will solve a rough classification is still very good when compared with that of no solution.

Further Research

This paper studies the ability of unsupervised classification algorithms on human mobility data to classify groups with a university campus. There are many practical avenues that the results of this study can be applied to and the options that further research can explore.

One area of further research is the exploration into improvements within the methodology. When assessing the areas on which improvement is needed, two parts stand out. The first looks to explore adding more features into feature set to analyse their effectiveness within mobility scenarios. This could include summarising behaviours linking to some of the parameters within the kth/campus data set such as x,y coordinates to summarise distances travelled. Another addition could look to build upon the accuracy of the base truth sets and use a more explorative process to indicate the best times within the system to collect that data.

Additionally, research can look to the possibilities of having more computational power. This would involve implementing the HoursOnCampus feature left out within this paper due to high computation times. With the use of more computationally powerful hardware, other research could be explored. This could refer to the extending of the timescale that the features would be summarising, this extension to could push the method defines week to a month or if the time could be devoted to it even a year. Such increases would rule out anomalies and aid more in-depth clustering of the scenarios.

Research could look to build upon the finding through using the classifications created and adding addition implementation to provide new uses for the data to inform analysis within human systems and to provide insight on how different groups use these spaces.

There are also opportunities for research to be extended with applications into new scenarios or locations that can utilise the same WiFi connections logs used within this paper. Such research could look to analyse more complex environments such at theme parks, where many more groups can be identified and this will test the clustering techniques versatility. Conversely, research could look to build upon the scenario already analysed moving to a similar scenario with minor changes in mobility between the known groups such as governments buildings.

Finally, as can be expected this research can't be considered perfect and such it is accompanied with flaws that still need to be solved. Implementations could look to analyse the methodology and then see where the solution has failed to provide the optimum solution and thus amendments could be made. I believe that although the ground truths used with in the paper provide good accuracy during evaluation but techniques could be evaluated to offer a better solution.

Conclusion

To conclude on the proposed solution, the final solution was evaluated to see how it had answered the initial Research Questions (RQs) and whether the hypothesis has accurately portrayed the data and the behaviour of the groups within.

When evaluating RQ1, the method and implementation have been able to both identify the known groups using and then classified them using a gaussian mixture unsupervised clustering technique. This RQ had the problem of being hard to evaluate due to the nature of the unlabelled data that was being used. The use of a ground truth provided enough to both confirm hypothesis predictions about the mobility of the classes and then using its footprint to evaluate the clustering classifications. Although no empirical score could be provided the nature of the task, ment that even though errors were certainly to be present on the whole a majority would have been classified correctly. But this could receive some more attention if further research was conducted.

When evaluating RQ2, research made obvious strides to meet the requirements that were placed by it. From this research the employed feature creation scheme and related features can be considered as a technique that aids is the summary of user movements as well as aiding the classification process. Research also showed a positive influence made from the processing techniques of PCA centring and MinMax normalisation on the data, although the features used should be considered carefully.

Although explicitly not spoken about within research RQ3 has been answered in some way with methodology decisions being made to aid methodology implementation over many different locations with that same foundation of human mobility. As found with the university the different classes are separated based on how sporadic their movements were, this then informed the clustering scheme. During the methodology decisions were made to generalise the solution, this included, using a broad scheme of features that were not specialised to the university scenario. The ground truths also helped aid this functionality with ground truths being universal in their application. As this is all theoretical further analysis on a novel data set within a new location would be needed to confirm its functionality.

RQ4 was the final research question within the study, and aimed at finding applications for the output produced within this paper and the benefit of producing such a classification. The benefits of such a classification can be linked to the further analysis that could happen with the addition information provided by the classification. This could include looking at the footfall behaviours of each class within the campus and then being able to provide analysis otherwise unachievable in this complex multiclass environment. This analysis could include fire safety planning for different groups within a system based on the differing movement behaviours, this could allow all for crowd control solutions to be formulated to analyse movement patterns for groups using the classification provided here. When looking for more commercial uses such classifications could be used unison with the WiFi logs to calculate workforce effectiveness. Finally classifications could provide needed for insight to provide increased efficiency of public transport allocation and asset management with these scenarios. This could be done with the tracking of groups to provide insight on when and where these assets are needed improving the running of these large areas.

Bibliography

Figure 1 – KTH Main Campus Map

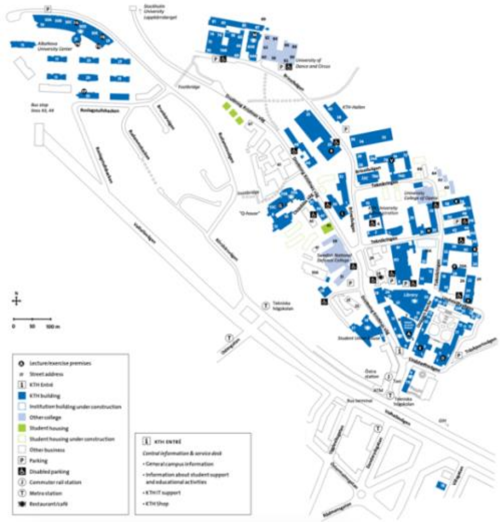

0 50 100 m
Lecture/exercise premises
Street address
KTH Entré
KTH building
Institution building under construction
Other college
Student housing
Student housing under construction
Other business
Parking
Disabled parking
Commuter rail station
Metro station
Restaurant/café
KTH ENTRÉ
Central information & service desk
• General campus information
• Information about student support and educational activities
• KTH IT support
• KTH Shop

Figure 2 – MaxMin Normalisation

$$x' = (x - \mu)/\sigma$$

where: μ = mean

σ = standard deviation

Figure 3 – System Flow Diagram

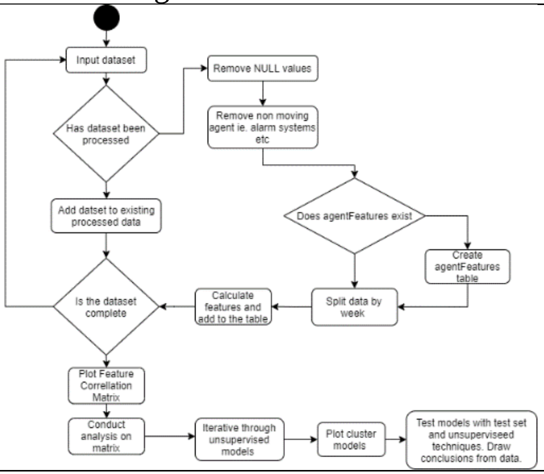
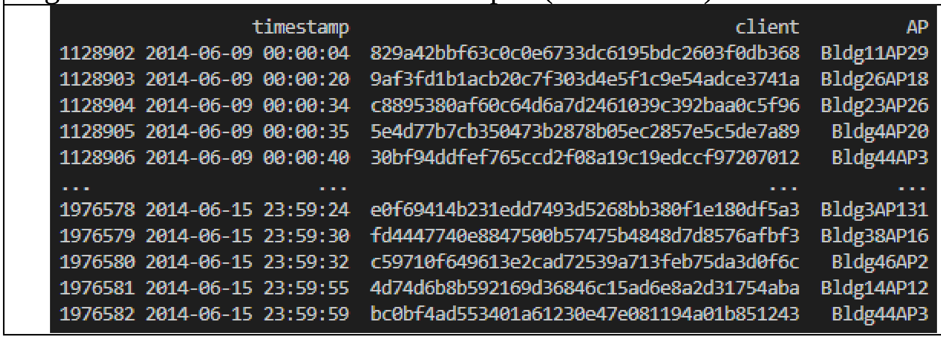
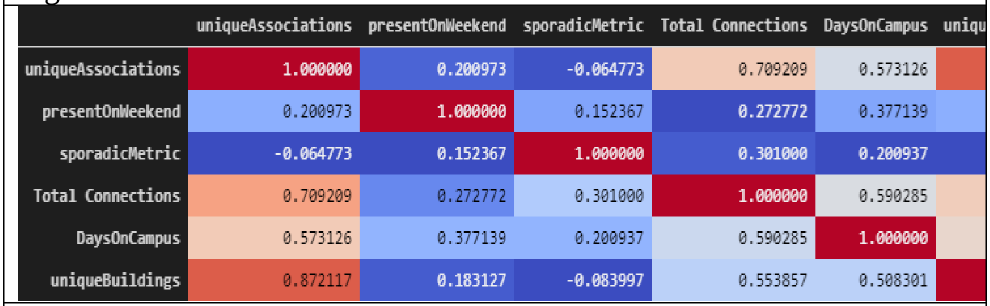

Input dataset
Has dataset been processed
Remove NULL values
Remove non moving agent ie. alarm systems etc
Add datset to existing processed data
Does agentFeatures exist
Create agentFeatures table
Is the dataset complete
Calculate features and add to the table
Split data by week
Plot Feature Correllation Matrix
Conduct analysis on matrix
Iterative through unsupervised models
Plot cluster models
Test models with test set and unsupervised techniques. Draw conclusions from data.

Figure 4 – Raw Data from KTH/campus (09-06-2014)



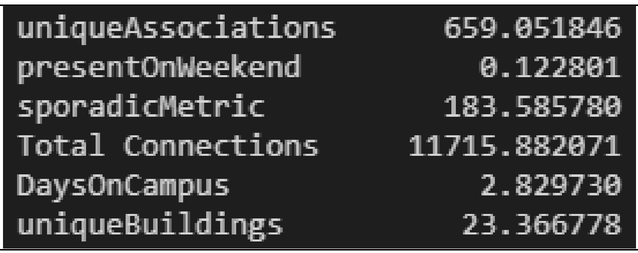
```
                  timestamp                            client        AP
1128902 2014-06-09 00:00:04  829a42bbf63c0c0e6733dc6195bdc2603f0db368  Bldg11AP29
1128903 2014-06-09 00:00:20  9af3fd1b1acb20c7f303d4e5f1c9e54adce3741a  Bldg26AP18
1128904 2014-06-09 00:00:34  c8895380af60c64d6a7d2461039c392baa0c5f96  Bldg23AP26
1128905 2014-06-09 00:00:35  5e4d77b7cb350473b2878b05ec2857e5c5de7a89   Bldg4AP20
1128906 2014-06-09 00:00:40  30bf94ddfef765ccd2f08a19c19edccf97207012   Bldg44AP3
...                     ...                                       ...         ...
1976578 2014-06-15 23:59:24  e0f69414b231edd7493d5268bb380f1e180df5a3  Bldg3AP131
1976579 2014-06-15 23:59:30  fd4447740e8847500b57475b4848d7d8576afbf3  Bldg38AP16
1976580 2014-06-15 23:59:32  c59710f649613e2cad72539a713feb75da3d0f6c   Bldg46AP2
1976581 2014-06-15 23:59:55  4d74d6b8b592169d36846c15ad6e8a2d31754aba  Bldg14AP12
1976582 2014-06-15 23:59:59  bc0bf4ad553401a61230e47e081194a01b851243   Bldg44AP3
```

Figure 5 – Data Correlation Matrix



	uniqueAssociations	presentOnWeekend	sporadicMetric	Total Connections	DaysOnCampus	uniqu
uniqueAssociations	1.000000	0.200973	-0.064773	0.709209	0.573126	
presentOnWeekend	0.200973	1.000000	0.152367	0.272772	0.377139	
sporadicMetric	-0.064773	0.152367	1.000000	0.301000	0.200937	
Total Connections	0.709209	0.272772	0.301000	1.000000	0.590285	
DaysOnCampus	0.573126	0.377139	0.200937	0.590285	1.000000	
uniqueBuildings	0.872117	0.183127	-0.083997	0.553857	0.508301	

Figure 6 – Variance of Feature Set



```
uniqueAssociations        659.051846
presentOnWeekend            0.122801
sporadicMetric            183.585780
Total Connections       11715.882071
DaysOnCampus                2.829730
uniqueBuildings            23.366778
```

Figure 7 – Initial Ground Truth (Non Moving)

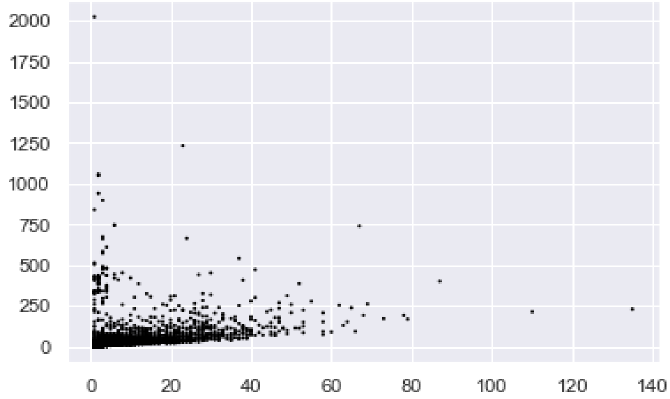


Figure 8 - Initial Ground Truth (Non Moving and Staff)

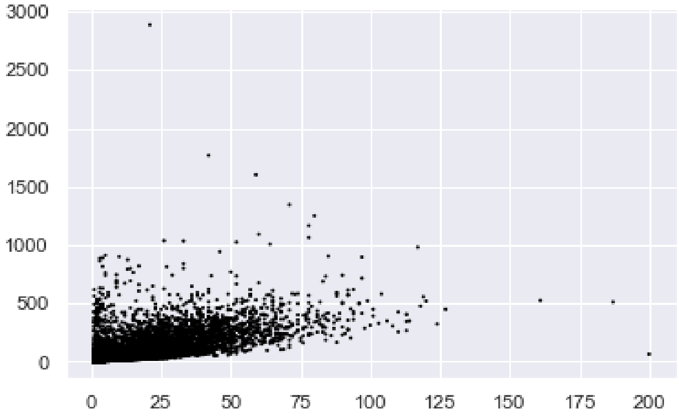


Figure 9 - Initial Ground Truth (Non Moving, Staff and Students)

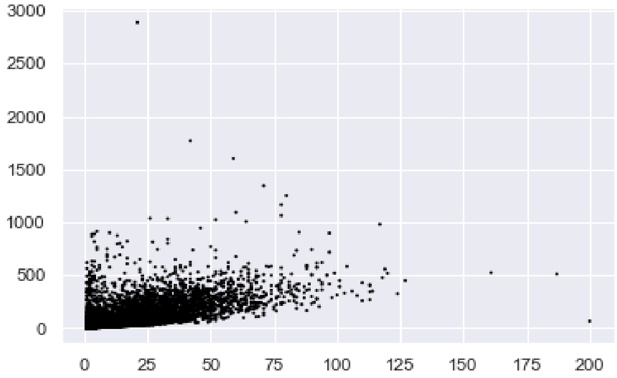


Figure 10 – Initial K-Means Classification: White (Non-Moving), Red (Staff), Pink (Students)

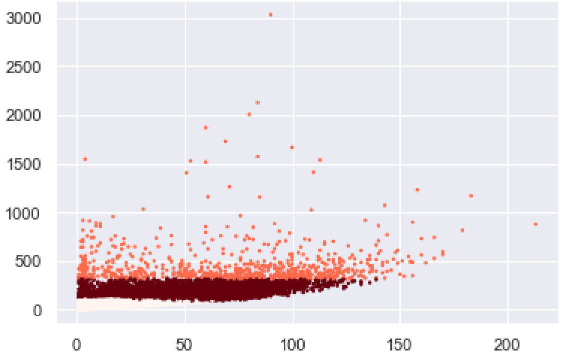


Figure 11 – Initial Gaussian Mixture Classification

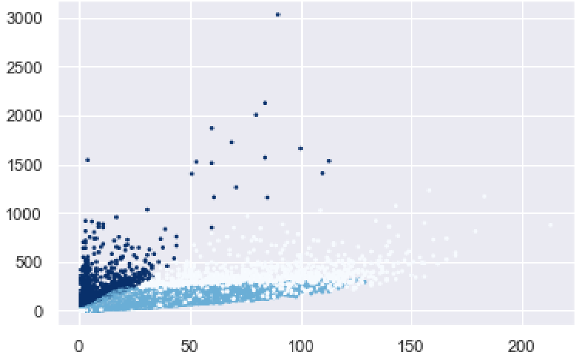


Figure 12 – Initial DBSCAN Classification

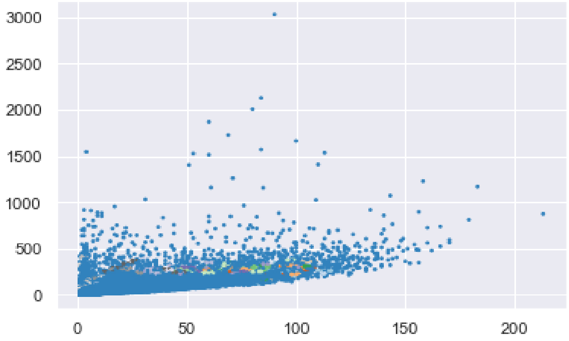


Figure 13 – PCA Ground Truth (Non Moving)

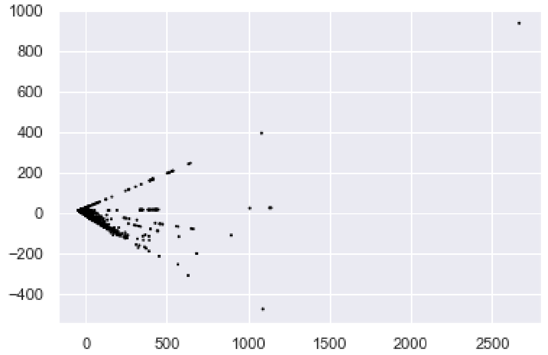


Figure 14 - PCA Ground Truth (Non Moving, Staff)

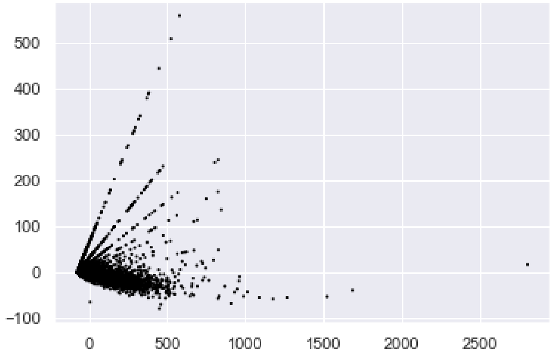


Figure 15 - PCA Ground Truth (Non Moving, Staff and Students)

Figure 16 - PCA Introduction K-Means

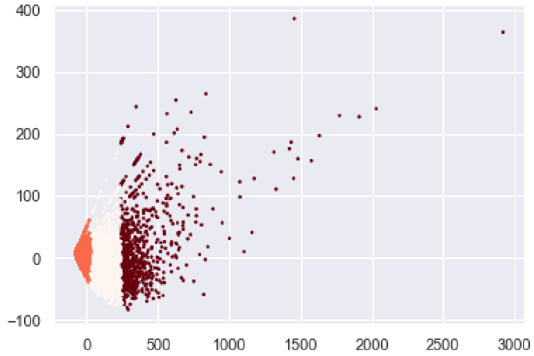


Figure 17 - PCA Introduction Gaussian Mixture Classification: Light Blue (Non-Moving), White (Staff) and Dark Blue (Students)

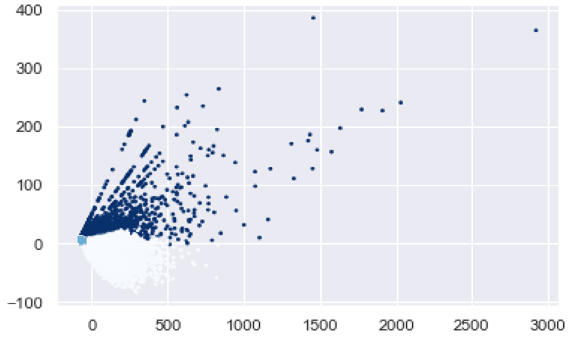


Figure 18 - PCA Introduction DBSCAN

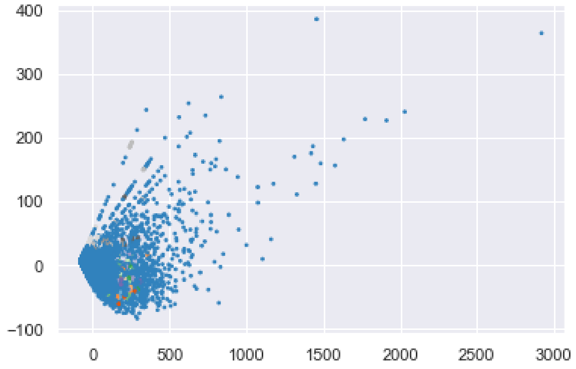


Figure 19 - PCA Introduction BIRCH

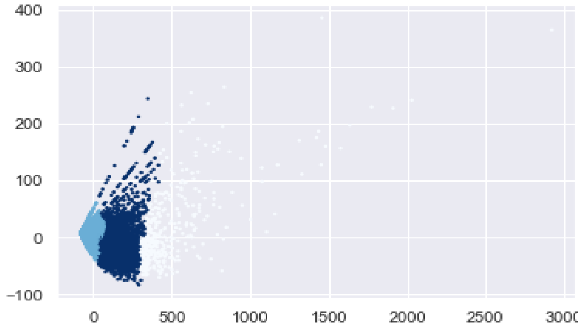
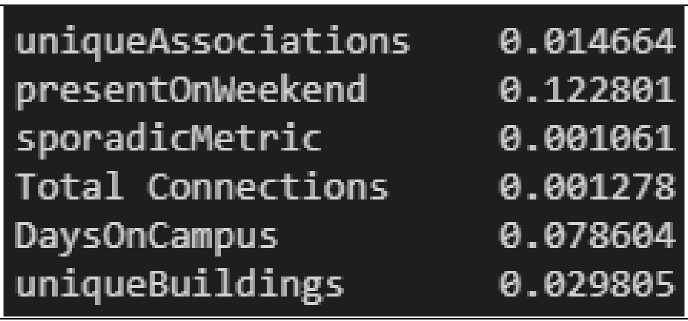


Figure 20 – Normalised Data Variance



```
uniqueAssociations     0.014664
presentOnWeekend       0.122801
sporadicMetric         0.001061
Total Connections      0.001278
DaysOnCampus           0.078604
uniqueBuildings        0.029805
```

Figure 21 – Data Shape After PCA and Normalisation

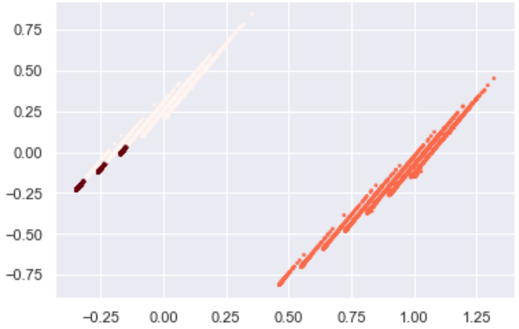


Figure 22 - Data Shape After PCA and Normalisation (Removal of presentOnWeekend)

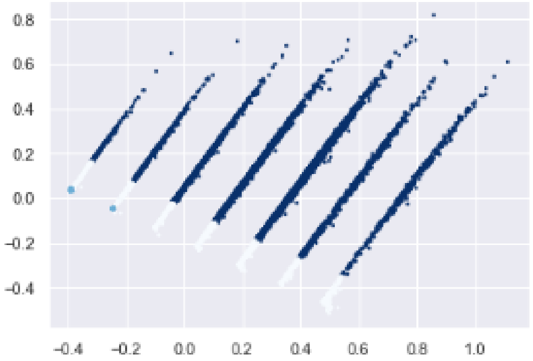


Figure 23 – PCA & Normalised Ground Truth (Non Moving, Staff and Students). Removal of DaysOnCampus and presentOnWeekend occurs in all figures after this one.

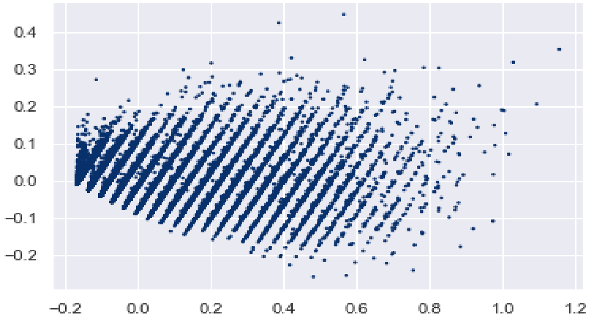


Figure 24 - PCA & Normalised Ground Truth (Non Moving, Staff)

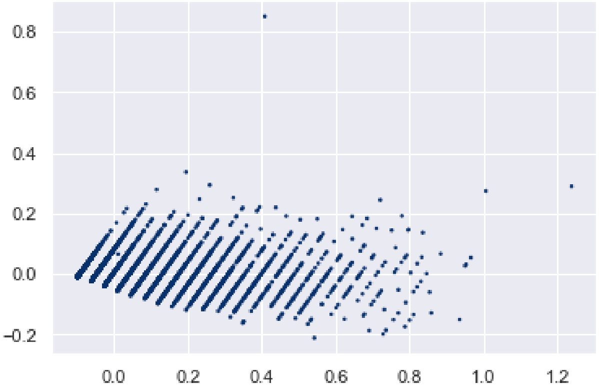


Figure 25 - PCA & Normalised Ground Truth (Non Moving)

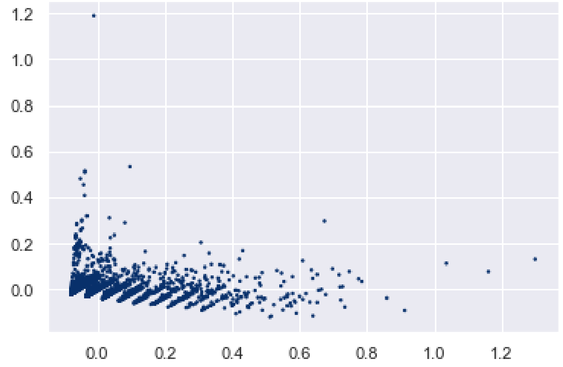


Figure 26 - PCA & Normalised K-Means

Figure 27 - PCA & Normalised Gaussian Mixture / Final Clustering Technique

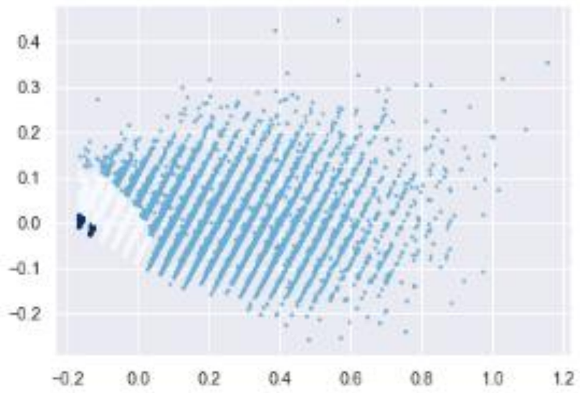
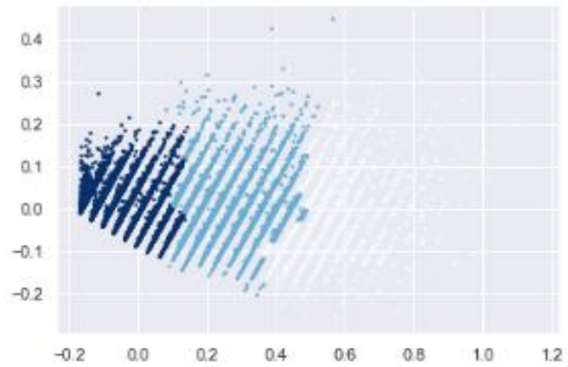


Figure 28 – PCA & Normalised BIRCH



Referencing

[1] M. M. Mackay and O. Weidlich, “Australian mobile phone lifestyle index,” *Australia: AIMIA-The Digital Industry Association of Australia,* 2014.

[2] G. Riccardo, B. Armando and R. Sandro, “Towards a statistical physics of human mobility,” *International Journal of Modern Physics C,* vol. 23, no. 09, p. 1250061, 2012.

[3] S. Phithakkitnukoon, Z. Smoreda and P. Oliver, “Socio-geography of human mobility: A study using longitudinal mobile phone data,” *PloS one,* vol. 7, no. 6, p. e39253, 2012.

[4] N. Keyfitz, “Individual mobility in a stationary population,” *Population studies,* vol. 27, no. 2, pp. 335--352, 1973.

[5] C. Song, Z. Qu and B. Nicholas, “Limits of predictability in human mobility,” *American Association for the Advancement of Science,* vol. 327, no. 5968, pp. 1018--1021, 2010.

[6] M. C. Gonzalez, C. A. Hidalgo and A.-L. Barabasi, “Understanding individual human mobility patterns,” *Nature* , vol. 453, no. 7196, pp. 779--782, 2008.

[7] X. Lu, E. Wetter, N. Bharti, A. J. Tatem and L. Bengtsson, “Approaching the limit of predictability in human mobility,” *Scientific reports,* vol. 3, no. 1, pp. 1--9, 2013.

[8] J. Feng, C. Rong, F. Sun, D. Guo and Y. Li, “PMF: A privacy-preserving human mobility prediction framework via federated learning,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies,* vol. 4, no. 1, pp. 1-21, 2020.

[9] I. Sirkeci and M. M. Yucesahin, “Coronavirus and migration: Analysis of human mobility and the spread of COVID-19,” *Migration Letters,* vol. 17, no. 2, pp. 379-398, 2020.

[10] A. Wesolowski, C. O. Buckee, L. Bengtsson, E. Wetter, X. Lu and A. J. Tatem, “Commentary: containing the Ebola outbreak-the potential and challenge of mobile network data,” *PLoS currents,* vol. 6, 2014.

[11] J. M. Marshall, S. L. Wu, S. S. Kiware, M. Ndhlovu, A. L. Ouedraogo, M. B. Toure, H. J. Sturrock, A. C. Ghani and N. M. Ferguson, “Mathematical models of human mobility of relevance to malaria transmission in Africa,” *Scientific reports.*

[12] R. O. loritun, T. B. Ouarda, S. Moturu, A. Madan, A. S. Pentland and I. Khayal, “Change in BMI accurately predicted by social exposure to acquaintances,” *PloS one,* vol. 8, no. 11, p. e79238, 2013.

[13] S. Shang, D. Guo, J. Liu and K. Liu, “Human mobility prediction and unobstructed route planning in public transport networks,” in *2014 IEEE 15th international conference on mobile data management,* 2014, pp. 43-48.

[14] Y. Zheng, Q. Li, Y. Chen, X. Xie and W.-Y. Ma, “Understanding mobility based on GPS data,” in *Proceedings of the 10th international conference on Ubiquitous computing}*, 2008, p. 321.

[15] K. Zhao, D. Khryashchev, J. Freire, C. Silva and H. Vo, “Predicting taxi demand at high spatial resolution: Approaching the limit of predictability,” in *IEEE international conference on Big data (big data),* IEEE, 2016, pp. 833-842.

[16] Y.-A. De Montjoye, C. A. Hidalgo and M. Verleysen, “Unique in the crowd: The privacy bounds of human mobility,” *Scientific reports,* vol. 3, no. 1, pp. 1-5, 2013.

[17] B. Tang, C. Jiang, H. He and Y. Guo, “Probabilistic human mobility model in indoor environment,” in *2016 International Joint Conference on Neural Networks (IJCNN),* IEEE, 2016, pp. 1601-1608.

[18] X. Li, D. Zhang, Q. Lv, J. Xiong, S. Li, Y. Zhang and H. Mei, “IndoTrack: Device-free indoor human tracking with commodity Wi-Fi,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies,* vol. 1, no. 3, pp. 1-22, 2017.

[19] P. Deville, C. Linard, S. Martin, M. Gilbert, F. R. Stevens, A. E. Gaughan, V. D. Blondel and A. J. Tatem, “Dynamic population mapping using mobile phone data,” *Proceedings of the National Academy of Sciences,* vol. 111, no. 45, pp. 15888-15893, 2014.

[20] R. Becker, K. Hanson, S. Isaacman, J. M. Loh and M. Martonosi, “Human mobility characterization from cellular network data,” *Communications of the ACM,* vol. 56, no. 1, pp. 74-82, 2013.

[21] A. Cangialosi, J. E. Monaly and S. C. Yang, “Leveraging RFID in hospitals: Patient life cycle and mobility perspectives,” *IEEE Communications Magazine,* vol. 45, no. 9, p. 9, 2007.

[22] G. Biczok, S. D. Martinez, T. Jelle and J. Krogstie, “Navigating MazeMap: indoor human mobility, spatio-logical ties and future potential,” in *2014 IEEE International Conference on Pervasive Computing and Communication Workshops*, IEEE, 2014, pp. 266-271.

[23] D. Brockmann, L. Hufnagel and T. Geisel, “The scaling laws of human travel,” *Nature*, vol. 439, no. 7075, pp. 462-465, 2006.

[24] E. Cho, S. A. Myers and J. Leskovec, “Friendship and mobility: user movement in location-based social networks,” in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining,* 2011, pp. 1082-1090.

[25] K. Zhao, S. Tarkoma, S. Liu and H. Vo, “Urban human mobility data mining: An overview,” in *2016 IEEE International Conference on Big Data (Big Data),* IEEE, 2016, pp. 1911--1920.

[26] X. a. N. N. Liu, X. Liu, H. Jin, J. Ou, L. Jiao and Y. Liu, “Characterizing mixed-use buildings based on multi-source big data,” *International Journal of Geographical Information Science,* vol. 32, no. 4, pp. 738-756, 2018.

[27] R. a. Z. K. Jurdak, J. Liu, M. AbouJaoude, M. Cameron and D. Newth, “Understanding human mobility from Twitter,” *PloS one,* vol. 10, no. 7, p. e0131469, 2015.

[28] J. S. Jia, X. Lu, Y. Yuan, G. Xu, J. Jia and N. A. Christakis, “Population flow drives spatio-temporal distribution of COVID-19 in China,” *Nature,* vol. 582, no. 7812, pp. 389-394, 2020.

[29] P. Sapiezynski, A. Stopczynski, R. Gatej and S. Lehmann, “Tracking human mobility using WiFi signals,” *PloS one,* vol. 10, no. 7, p. e0130824, 2015.

[30] R. Pellungrini, L. Pappalardo, F. Pratesi and A. Monreale, “A data mining approach to assess privacy risk in human mobility data,” *ACM Transactions on Intelligent Systems and Technology (TIST),* vol. 9, no. 3, pp. 1-27, 2017.

[31] F. Pratesi, A. Monreale, R. Trasarti, F. Giannotti, D. Pedreschi and T. Yanagihara, A System for Assessing Privacy Risk versus Quality in Data Sharing, Pisa, Italy, 2016.

[32] D. Soper, “Is human mobility tracking a good idea,” *Communications of the ACM,* vol. 55, no. 4, pp. 35-37, 2012.

[33] D. College, “Crawdad - A Community Resource for Archiving Wireless Data,” [Online]. Available: https://crawdad.org/. [Accessed 11 December 2020].

[34] “Dartmouth College,” [Online]. Available: https://home.dartmouth.edu/. [Accessed 11 December 2020].

[35] “Crawdad Data License,” [Online]. Available: https://crawdad.org/data-license-agreement.html. [Accessed 12 December 2020].

[36] L. a. F. V. a. K. G. Pajevic, “Revisiting the modeling of user association patterns in a university wireless network,” in *2018 IEEE Wireless Communications and Networking Conference (WCNC),* IEEE, 2018, pp. 1-6.

[37] M. W. Traunmueller, N. Johnson, A. Malik and C. E. Kontokosta, “Digital footprints: Using WiFi probe and locational data to analyze human mobility trajectories in cities,” *Computers, Environment and Urban Systems,* vol. 72, pp. 4-12, 2018.

[38] J. Irani, N. Pise and M. Phatak, “Clustering techniques and the similarity measures used in clustering: a survey,” *International journal of computer applications,* vol. 134, no. 7, pp. 9-14, 2016.

[39] C. Castellano, S. Fortunato and V. Loreto, “Statistical physics of social dynamics,” *Reviews of modern physics,* vol. 81, no. 2, p. 591, 2009.

[40] A. D. Nguyen, P. Senac, V. Ramiro and M. Diaz.

[41] S. Hasan, C. M. Schneider, S. V. Ukkusuri and M. C. Gonz'alez, “Spatiotemporal patterns of urban human mobility,” *Journal of Statistical Physics,* vol. 151, no. 1, pp. 304-318, 2013.

[42] C. Cattuto, W. Van den Broeck, A. Barrat, V. Colizza, J.-F. Pinton and A. Vespignani, “Dynamics of person-to-person interactions from

distributed RFID sensor networks," *PloS one,* vol. 5, no. 7, p. e11596, 2010.