UNIVERSITY OF NOTTINGHAM

Managing the Uncertainty of Occupant Behaviour for Building Energy Evaluation and Management

Sophie Naylor

2018

ABSTRACT

The influence of building occupancy and user behaviour on energy usage has been identified as a source of uncertainty in current understanding of operational buildings, and yet it is rarely directly monitored. Gathering data on the occupancy of buildings in use is essential to improve understanding of how energy is used relative to the actual energy requirements of building users.

This thesis covers the application of occupancy measurement and processing techniques in order to address the gap in knowledge around the contextual understanding of how occupants' changing use of a building affects this building's optimum energy demand in real time. Through targeted studies of running buildings, it was found that typical current occupancy measurement techniques do not provide sufficient context to make energy management decisions. Useable occupancy information must be interpreted from raw data sources to provide benefit: in particular, many slower response systems need information for pre-emptive control to be effective and deliver comfort conditions efficiently, an issue that is highlighted in existing research. Systems utilising novel technologies were developed and tested, targeted at the detection and localisation of occupants' personal mobile devices, making opportunistic use of the existing hardware carried by most building occupants. It was found that while these systems had the potential for accurate localisation of occupants, this was dependent on personal hardware and physical factors affecting signal strength. Data from these sources was also used alongside environmental data measurements in novel algorithms to combine sensor data into a localised estimation of occupancy rates and to estimate near-future changes in occupancy rate, calculating the level of confidence in this prediction. The developed sensor combination model showed that a selected combination of sensors could provide more information than any single data source, but that the physical characteristics and use patterns of the monitored space can affect how sensors respond, meaning a generic model to interpret data from multiple spaces was not feasible. The predictive model showed that a trained model could provide a better prediction of near-future occupancy than the typically assumed fixed schedule, up to an average of approximately two hours.

The systems developed in this work were designed to facilitate the proactive control of buildings services, with particular value for slower-response systems such as heating and ventilation. With the application of appropriate control logic, the systems developed can be used to allow for greater energy savings during low or non-occupied periods, while also being more robust to changing occupant patterns and behaviours.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

ACKNOWLEDGEMENTS

With thanks to Prof Mark Gillott, Dr James Pinchin and Dr Ed Cooper for their guidance and supervision through this work.

Many thanks to those in the Engineering Excellence Group at Laing O'Rourke, in particular Mr Tom Lau and Dr Graham Herries for their supervision and valuable insights into industry perspectives throughout the project.

I would also like to acknowledge the contributions from parties involved in the collection of data for this work, including Cristian Becerra Monroy for his patience in setting up an EnOcean data collector in the Mark Group House, the team at Beckhoff for their efforts in the setup and hosting of the data platform, the staff of Laing O'Rourke who kindly provided access control data for their offices and Pressac Communications for providing sensors.

Funding for this project was provided by the EPSRC and Laing O'Rourke.

A final thanks to the family and friends who have personally supported me through this work.

1 INTRODUCTION

1.1 General Introduction

One of the major sources of energy demand in the UK is the built environment. Recent data [1],[2] shows that over 40% of the UK's energy consumption occurs in buildings. More than 80% of energy used over a building's life cycle occurs during its operation [3]. As the built environment is the largest single energy consumer in the UK, it is essential that effort is made to reduce the amount of fossil fuel energy used by buildings – both by diverting some demand to renewable resources and, more importantly, by addressing inefficiencies in the way that buildings are operated in order to lower the demand for energy during building use.

Data collected during building operation typically shows a significant 'performance gap' between designed and actual energy use in buildings, with this trend consistent across multiple sectors [4]. Discrepancies between predicted and real building performance are caused by both an underestimation of predicted values for reasonable building use at the design stage, and a greater than expected use of resources during the running building's life [5]. Among other factors, such as variations in the delivered performance of building fabric and assumptions made in design-stage modelling tools, both sides of the discrepancy are affected by the "inability of current modelling methods to represent realistic use and operation of buildings" [6]. Occupant behaviour is one of the factors contributing to excessive energy use during building operation, alongside the effectiveness of building services controls to meet occupant energy demands. Uncertainties around the way that occupants actually use buildings further contribute to the inability to meet changing occupant energy requirements in buildings services control/energy management systems.

Many of the routes to reducing the performance gap therefore rely on a way to effectively reduce the current level of uncertainty around building occupants and their behaviours. In particular, building control/energy management systems seeking to reduce energy waste while maintaining or improving occupant comfort must have a way to gauge realistic occupant energy demand in real time. However, the field of detailed exploration into occupant behaviour is relatively young, meaning that there is a lack of in-use data for buildings in the UK [7]. In particular, long-term data on

occupancy within working buildings is rarely collected. In the commercial sector, organisations such as CarbonBuzz are working to address the general lack of in-use data by inviting case studies comparing the design and actual energy use of working buildings, although explicit measurement of occupancy rates is not required [8]. The CarbonBuzz resource states that there is a disconnect in the data available to building designers and users, making it difficult to fully understand where building design and operation is failing. Current commercial building controls systems often fail to account for occupants beyond simple scheduling, leading to wasteful conditioning of spaces or control of systems where occupants are not present. In general, there is a clearly identified need for a more widely applicable and comprehensive way to gather relevant occupant data in order to better inform energy management decisions.

This thesis seeks to address this need by occupancy measurement techniques targeted to improve the contextual understanding of how occupants' changing use of a building affects this building's energy demand in real time.

1.2 Aim

The aim of this work is to develop novel systems and algorithms for the measurement and prediction of localised building occupancy rates. These systems are aimed towards integration with building energy management systems, to tailor energy used in buildings more closely to the actual needs and behaviours of occupants.

1.3 Objectives

The scope of this research was divided into several more specific objectives:

1. To assess the current state of occupant-centred Building Energy Management Systems and indoor occupancy measurement technologies.
2. To understand existing approaches to applied indoor occupancy measurement and highlight the gap in existing knowledge.
3. To develop a specification for novel indoor occupancy sensing and prediction methods suitable for use in building controls systems.
4. To develop and test a method for measuring localised indoor occupancy rates
5. To develop and test a method for predicting short-term future localised occupancy rates based on historical data, with sensitivity to the unavoidable uncertainties involved in behavioural prediction.

1.4 Research Methodology

In order to address the objectives set out for this work, the following research methodology was applied.

Objective	Method(s)
1	- Review in detail the market and current state of technology used in commercial Building Energy Management Systems
2	- Review and compare existing technologies and methods used for indoor occupancy measurement - Review and compare existing approaches to applying occupant data in a Building Energy Management context
3	- Assess existing relationships between occupancy and energy use through case studies on a range of building types - Identify through case studies where uncertainties around occupant patterns and behaviours can be reduced through more targeted data collection - Define and develop data collection methods targeted towards identified needs - Combine knowledge from review work, case studies and data collection development into a proposal for an improved occupancy measurement framework
4	- Select appropriate methods for the proposed measurement of localised occupancy rates - Develop and test methods for localised occupancy rate measurement
5	- Select appropriate methods for the proposed short-term future occupancy prediction - Develop and test methods for short-term occupancy prediction

1.5 Contributions to Knowledge

The work presented in this thesis covers the development of novel systems and algorithms for the detection and future prediction of localised occupancy rates. This is achieved by the following streams of innovation. Occupancy data is gathered through multiple sources, including regular environmental sensor readings of motion and CO_2 level and the presence and location of personal mobile devices using Wi-Fi and Bluetooth-Low-Energy technologies. The Wi-Fi detection involves the hardware and software development of a novel low-cost setup to gather and process presence data, while the Bluetooth detection uses zoned beacon placement to gather relevant location data without excessive data processing and storage requirements. Data collected is processed to build models for the interpretation of raw sensor data into useable measures of occupancy: a task that is addressed through a proposed modular approach to parsing occupant information. Neural network models are developed and tested for inferring local occupancy rates from sensor data, and for predictive probabilistic modelling of short-term occupancy informed by current measured occupancy. The

level of uncertainty in the predictive output is explicitly quantified, allowing more sensitive treatment of the model output in high-uncertainty situations. This system is designed to inform proactive building control through BEMS, allowing for more effective use of slow-response building services such as heating and ventilation.

The work also combines broader studies of occupancy and energy in a range of building types, assembling a database of in-use performance to allow informed and direct targeting of identified energy wastes in relation to building occupancy/behaviour. This analysis of in-use data could be valuable in industry, where the delivery of buildings that perform to specification is a priority.

1.6 Thesis Structure

Chapter 2 introduces a basis of the current state of buildings in the UK and current practices in building automation, providing a breakdown of building automation terms used in the industry. The influence of building occupants on indoor energy demand is also examined through a review of studies in this area.

Chapter 3 presents a review and cross-analysis of existing work into sensing technologies for building occupants, providing detail on the quality of occupant information that can be obtained with varying sensor types and methods. Existing studies into how varying grades of occupant data can be used in building controls are also assessed, with consideration for how the intended application shapes the control responses to occupant data.

In Chapter 4, several case studies are made using data collected from live buildings to assess in more detail the relationship between occupancy and energy use across a range of building types, and to evaluate the available methods for occupant data collection.

Chapter 5 introduces the major data source for this study – a small office building based in Nottingham, UK, fitted with a range of localised environmental sensors. This chapter also provides details on the development and testing of Wi-Fi and Bluetooth-based systems to detect building occupants using personal mobile devices.

Chapter 6 covers the selection and application of appropriate machine learning techniques to infer the local number of occupants from raw sensor data collected in

the small office building testbed. The value of each of the sensor types available is assessed in its usefulness for this task, and an optimised model proposed.

Chapter 7 covers the development of a machine learning model to predict future local occupancy rates based on recent past behaviours. Using generated long-term datasets with varying underlying occupancy patterns, the model is tested against simpler prediction heuristics to assess its value in differing situations. A final model is proposed to make use of the uncertainty level in the predicted output to defer to a safer strategy when uncertainty is high.

Chapter 8 provides final conclusions and discussion of future work.

2 REVIEW OF EXISTING MARKET FOR BUILDING CONTROLS

2.1 Introduction

To provide a basis for making a better system to respond to occupant needs in buildings, a review was made into the current state of building energy use, current building energy management systems practise and the influence of occupants and occupant behaviours in building energy demand. This chapter presents the findings of this review.

2.2 UK Building Stock

This section focuses on current energy issues in the existing building stock in the UK. As mentioned in the introduction, buildings represent over 40% of the UK's overall energy demand, with split approximately two thirds domestic and one third non-domestic, as illustrated in Figure 2-1.

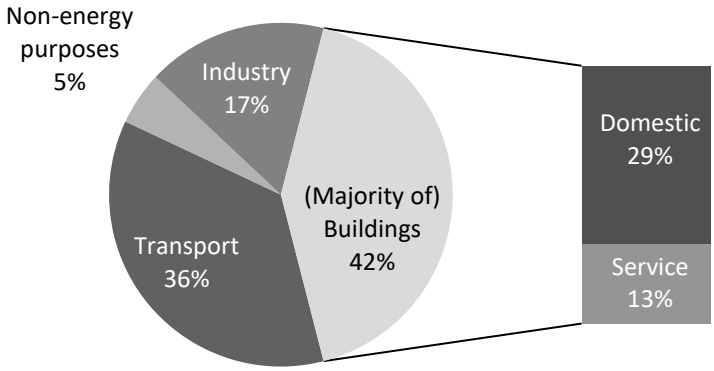


Figure 2-1 – Energy Use in the UK by Sector (2012) [2]

2.2.1 Domestic Buildings

Domestic buildings made up 29% of the UK total energy consumption in 2012 [9]. As Figure 2-2 shows, the majority of this energy is used for space heating and domestic hot water. While domestic hot water use is largely behavioural, domestic space heating systems often include some degree of automated control.

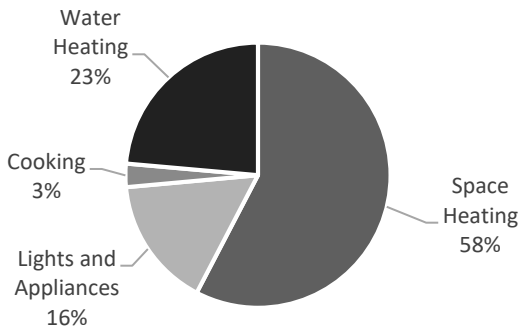


Figure 2-2 - UK Domestic Energy Demand by Fuel and End Use [10]

There is significant scope to improve the energy efficiency of residences in the UK. Given that the majority of houses show poor material performance (see Figure 2-3) [11][12], it is likely that those homes wasting the most energy would benefit the most from simple material retrofits such as adding insulation. In terms of costs to the building owner, cavity wall insulation is typically priced in order of £100 [13], while home automation can cost in the order of £1000-£10000 depending on complexity [14]. This suggests that, for the worst performing homes, introducing advanced controls is not the most viable immediate option. However, recent trends of interest in home automation and integrated 'smart home' products show that there is still considerable scope for the implementation of advanced controls in domestic buildings, with significant energy saving possible.

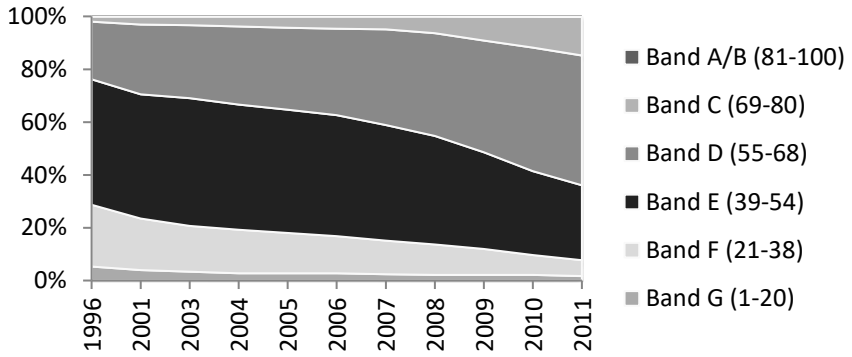


Figure 2-3 - UK Dwelling Energy Efficiency Rating 1996-2011 [11]

2.2.2 Non-Domestic Buildings

The service sector (including public buildings, workplaces, schools, hotels etc.) accounts for around 13% of the total UK energy consumption [15], while industrial buildings use 17% [16]. The majority of industrial energy consumption is used for processes and thus not the focus of this study.

The performance of non-domestic service buildings in the UK is highly varied depending on building purpose. Figure 2-4 summarises the end-use energy consumption for various building types. It can be seen that the retail and education sub-sectors consume the highest levels of energy, mostly due to space heating and lighting.

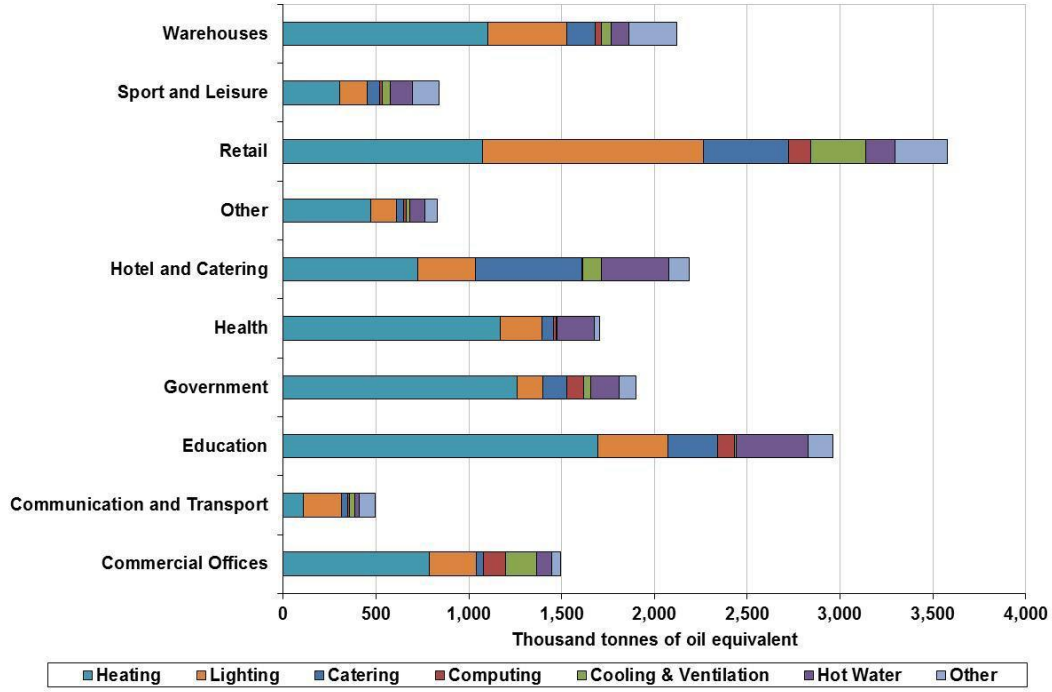


Figure 2-4 - Service sector energy consumption by end use and sub-sector, UK (2012) [15]

2.2.3 Discussion

It can be seen that the energy use during operation of buildings across sectors can be highly variable and is spent largely on creating thermal conditions that satisfy the comfort of building occupants. However, there is currently a significant lack of understanding into how exactly energy consumption within a building relates to the behaviours, comfort and response of occupants. This discrepancy can be addressed by the implementation of more comprehensive sensing of buildings during operation and more sophisticated controls.

2.3 Building Energy Management Systems

The automated control of buildings encompasses an array of different technologies and is described using many different terms, depending on the context, area of application, system manufacturer etc. This section explains the implications and overlap of various terms in the building management field.

Buildings adopting automation technology:

- Green Building – a building that employs technologies to save energy and minimise negative effects to the environment. Does not directly imply automation, although many green buildings will use some form of automated system as part of their strategy.
- Smart/Intelligent Building – definition varies between sources and professional bodies [17], but generally implies a building with some form of automated control of services to create a more comfortable or environmentally friendly space [18].
- Sentient Building – more rarely used, implies a step beyond 'intelligence'; a building with high levels of data collection and automation, possibly using a continuously updated working model of the whole building [19].

Residential buildings with automation technology:

- Home automation – the field of building services automation when applied specifically to a domestic building.
- Domotics – 'a contraction of "domus", meaning "home", and the words informatics, telematics and robotics' [20], refers to home automation.
- Smart home – a house that uses internet-connected 'smart' devices for automation of various functions, including building services [21].

Names for the automation system:

- BMS/BAS(Building Management/Automation System) – these terms are used interchangeably [22] and can refer to any building services system using automation. Most commonly used by manufacturers of automation equipment.
- BACS (Building Automation and Control System) – less common term for BAS/BMS [23].
- BEMS (Building Energy Management System) – encompasses building automation and energy priorities, with the implication that data is more intelligently analysed than in a purely automated system. 'Active BEMS' can be used to refer to a BMS with particular focus on energy saving or a sub-section of the greater BMS dedicated to energy saving. 'Passive BEMS' refers

to a piece of software to analyse energy use from BMS or sensor data and is not directly in control of the building services [24].

- EMCS/EMS(Energy Management and Control System/Energy Management System) – occasionally used to refer to automated control systems with focus on energy saving, although used in the wider industry to refer to broader energy demand and power quality controls [25].

Related terms:

- DCV (Demand Controlled Ventilation) – HVAC strategy to supply only the heating/ventilation/cooling needed by the actual number of occupants in a space [26].
- DDC (Direct Digital Control) – general term for automated control by a digital device [27].
- EIS (Energy Information System) – software used for data collection, visualisation and analysis, not necessarily control [28].

In this project, the term 'BEMS' is used throughout for simplicity. This reflects the priority towards energy saving and potential for more software-based solutions better than using the terms 'BMS' or 'BAS'.

2.3.1 Structure of BEMS

The hardware of a BEMS is typically divided into three layers, as shown in Figure 2-5. Each layer has its own functions and will pass data to other layers as appropriate.

1. **Management Layer –** This layer contains the supervisory software to control the BEMS and provides an interface for the user. Features may include automated and manual central control and energy feedback. In hardware terms, here the primary bus connects logic controllers, workstation terminals and web servers [29].
2. **Automation Layer –** In hardware terms, here the secondary bus connects to the controllers for the major components of the building services system, such as lighting, boiler, central plant and ventilation controllers. Localised controllers called 'outstations' make up the majority of this layer [30].
3. **Field Layer –** Provides the system's physical interface with the operation of the building. This includes a network of sensors (temperature, CO_2, motion,

electricity meters etc.) and actuator devices (control valves, light dimmers, automated switches etc.) throughout the building.

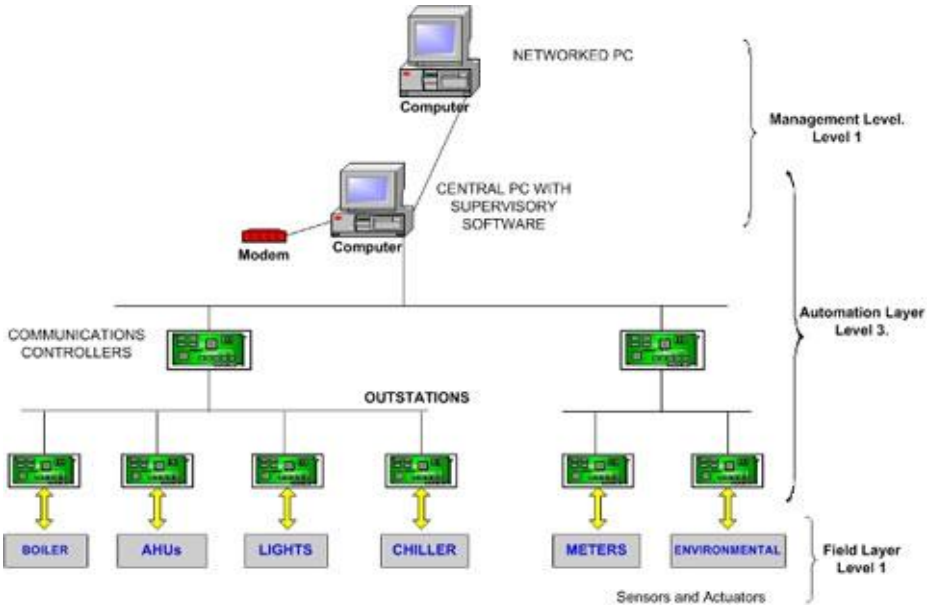


Figure 2-5 – Building Management System structure [31]

Most current energy management systems are not equipped with the field-level sensing equipment to adequately account for local occupant energy demands. Controls for each type of system are often isolated from each other, with poor interoperability. This makes the opportunistic use of diverse sensor types more difficult if some systems and their sensors are not available for decision making processes. At the higher level, management strategies must be developed that are more sensitive to the dynamic energy requirements of occupants and their energy behaviours. In order to do so, the following improvements must be made standard:

- Implementation of appropriate sensing equipment to detect localised changes in occupant energy demand.
- Processing applied to interpret sensor data into a defined measure of the occupancy data of interest: be that local occupant presence, location, number of occupants etc.
- Logical processing of occupant data to inform demand-driven control decisions.

2.4 BEMS Manufacturers and Market

BEMS and other automation in buildings play a significant role in the vision of 'ubiquitous computing' or 'ambient intelligence' – the concept that, in the future, intelligence will be built into the environment around people in order to learn the context of human behaviours and improve function automatically [32]. As such, the area of energy and comfort management has received significant research interest over the last decade [33].

Despite this vision, an estimated 90% of current HVAC control systems do not run optimally [34]. This shows the need for an improvement in the way that controls are designed and implemented.

2.4.1 Major Forces in the BEMS Industry

While a large range of hardware and software companies manufacture components for commercial BEMS and building automation, around 70% of business in the field is shared between 5 companies [35]:

- Honeywell [36] – A US-based conglomerate company with both the main company and several subsidiaries providing building automation worldwide. Subsidiaries include: Trend Controls (UK); Alerton (US); Novar (US) and Tridium [37].
- Johnson Controls [38] – US-based company specializing in control hardware and software.
- Schneider Electric [39] – produces automation and energy management components, based in France.
- Siemens [40] – German conglomerate engineering company, produces building automation hardware and software.
- United Technologies Corp.[41] – conglomerate company based in the US. Prominent in both aerospace and building systems industries.

The above large companies excel at producing the hardware and network components required for building automation, but recent industry reports hint towards the importance of the software and data analysis side of the business in the near future [24]. It is possible that smaller companies with an expertise in software development could become more prominent in the future of BEMS.

Where commercial systems remain slower to react to recent trends towards more occupant centric data-driven control decisions, the domestic market for home automation has seen a significant increase in interest during recent years, with a wide range of products and services released specifically for the improvement of domestic building control. Table 2-1 summarises the capability of several of the most popular 'smart' domestic heating control systems. It can be seen that many major controllers are adopting learning algorithms and occupancy-responsive technology, facilitated by the wider availability of easily installed wireless sensors. Trends towards occupant-responsive systems and integration with personal devices for remote control can also be seen in popular home automation management hubs and software. These 'central hub' solutions typically offer a more open system compatible with third party hardware using multiple communication protocols [42]. The vision of the 'internet of things' connecting a comprehensive network of devices in domestic settings has motivated several companies to develop domestic-level data collection on occupants and their interactions with home technologies and appliances.

Table 2-1 – Comparison of major 'smart' domestic heating controls packages available in the UK

	Hive Active Heating 2 (2016) [43]	**Nest v3 (UK) (2017) [44]**	**Heat Genius (2016) [45]**	**Tado (2016) [46]**	**Honeywell Evohome (2015) [47]**
Multi-zone control	Up to 3 zones if boiler allows, no TRVs	Per thermostat if boiler allows, no TRVs	♦	Per thermostat, no TRVs	♦
Remote Control	♦	♦	♦	♦	♦
User Motion Sensing		♦	♦		
User Geolocation	Limited - prompts manual alterations	♦		♦	
Learning heat response		♦	♦	♦	♦
Weather data use		♦	♦	♦	
Additional features	• Security Integration	• Self-learning schedules* • Security Integration	• Modular • Further home automation	• Distance dependent temperature setback	• High control granularity

* Reviews indicate effectiveness of learning can vary **[48]**

The greater sophistication of domestic systems can be difficult to apply to larger, more complex commercial systems, where a high number of occupants may cohabit in a space and obtaining occupancy/behavioural data becomes more complex.

2.5 Evidence for the Importance of Occupant Data

Data collected during building operation typically shows a significant difference between designed and actual energy use in buildings. This is clearly demonstrated in

Figure 2-6, which highlights the higher than expected use of both heat and electrical energy in various commercial building types. There is a greater difference in electricity use than heat, which likely reflects the fact that commercial building occupants have more control over electrical services (lighting, computers etc.) than heating controls. As such, it is easier for electricity to drift further away from design values.

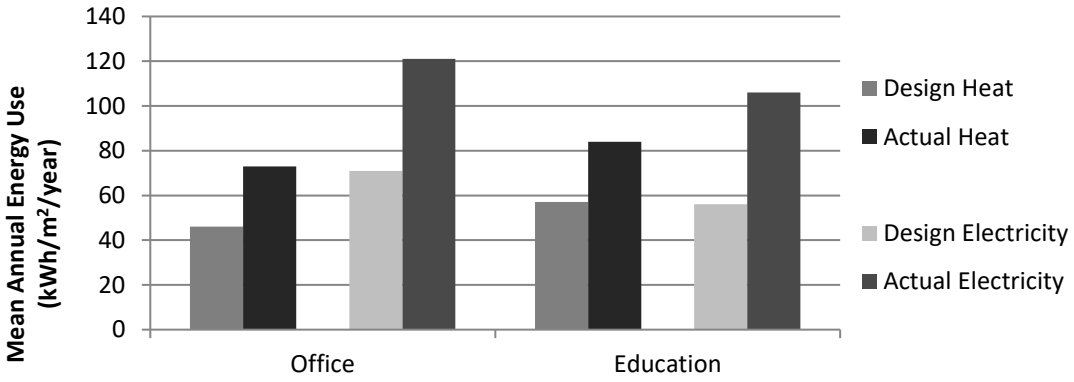


Figure 2-6 - Predicted Versus Actual Energy Use in Commercial Buildings [4]

Several studies have been conducted to assess the contribution of occupancy/occupant behaviour towards demand and final energy use in buildings. This is typically achieved through simulation or observation of a small set of real-world buildings. The findings of such studies are summarised below.

2.5.1 Impact of Occupant Behaviour in Commercial Buildings

Firstly, research through simulation typically tries to verify how much energy use changes with varying occupancy, allowing buildings' sensitivity to actions made by the occupant to be quantified.

For example, early sensitivity analysis of a school building [49] showed an energy use variation of as much of 150% within the bounds of "typical" occupant behaviour. It should be noted that the methods used involved static occupancy schedules, meaning that only extreme values of behaviour could be tested and a variation this high is unlikely in real building use.

Simulation of a medical building [50] has shown an overall energy variation of 30% when comparing the use of real occupancy patterns in building simulation to the same building with standard occupancy templates from simulation software. The simulated

person count and electrical equipment use were altered by up to ±68%. This verifies that energy use can be highly sensitive to occupant actions.

Azar and Menassa state that the 30-100% discrepancy between simulated and actual building energy use data "can mainly be attributed to misunderstanding and underestimating the important role that the occupants' energy use characteristics play in determining energy consumption levels" [51]. In their study, an agent-based model was used to account for changes in attitude towards energy and word-of-mouth effects between building occupants. Different input influence levels changed energy use in the tested building by as much as 25%, given the same behavioural starting point. This suggests that influence and change in behaviour over time is a factor that should be considered in building modelling. However, further sensitivity analysis on other building locations and sizes show that, while still significant, the relative impact of occupant behaviour changes with different building contexts [3].

Secondly, studies of control systems and real buildings quantify the impacts of current building services systems' response to changing occupancy. Simulation will typically assume the building services respond properly to changing demand from occupants, which is not the case in practice. Typically, a poor response to occupant presence and behaviour shows systems wasting energy by running when occupants are not present.

Martani et al demonstrate that, in real application, building services do not always follow actual occupant presence patterns [52]. Wi-Fi connections were used as an occupancy counting device in an educational building. Electricity use showed strong correlation with occupancy levels, while HVAC energy use did not. This shows the poor response of some services systems to actual occupancy. The authors observed that "large common areas, such as studios, may be used by one person or a large number of people often with no alteration in the amount of energy supplied to the space".

Masoso and Grobler's study of commercial buildings in a hot, dry climate showed more energy used during non-working hours than working hours [53]. This suggests that energy use is not properly linked to periods of occupancy and, once again, shows the need for more occupancy-centric control systems.

2.5.2 Impact of Occupant Behaviour in Domestic Buildings

A study of dwellings in Japan [54] took an alternative approach to quantifying occupant impact on energy: by eliminating other factors from a large data set until differences in behaviour were the only remaining factors to account for energy differences between buildings. This was done by grouping buildings with similar size, climate, number of residents etc. The study found that HVAC and most electrical equipment loads were heavily influenced by occupant behaviour. Hot water and refrigeration were not highly influenced by occupant behaviour.

Observational study of dwellings in Northern Ireland have shown that houses with similar characteristics can have different electricity use profiles, depending on the number of residents, patterns of occupancy and socio-economic factors affecting energy behaviours [55].

A study of PassiveHaus sensitivity showed that occupant presence patterns did not affect energy use as significantly as expected [56]. This may be symptomatic of poor control systems, as with the studies of commercial buildings in Section 2.5.1. It should be noted that occupant-controlled behaviours such as appliance use, set point temperatures and airflow behaviour were shown to have a significant impact on overall energy demand.

2.5.3 Discussion

Most research looking specifically at behavioural impact on building energy use shows that there is significant potential for energy saving through greater understanding of building use. Some studies offered insight into particular areas currently affected by changing behaviour: use of electrical appliances, lighting etc. HVAC loads are often not correlated to occupant presence or behaviour, suggesting there is scope to reduce wasted energy by making the energy supplied more closely match the varying demand from occupants.

Studies have noted considerable differences in motivation for energy-related behaviour in commercial and domestic settings (energy bill responsibility, privacy, social factors, typical activities etc.), suggesting that approaches to understand or adjust behaviours must be approached sensitively to their context [57]. Further to this observation, studies that considered several building types or settings show that occupant behavioural patterns can show significant variation between building types

[58], suggesting that generic solutions to behavioural modelling and control problems will not provide adequate performance for all buildings.

The above studies suggest that occupancy data can be used to effect change in a number of ways (more occupant-centric controls, feedback to occupants, informing the design process etc.) This work is focussed on improvements to the occupancy measurement systems providing information to building controls and energy management.

2.6 Conclusions

This review covered the existing state of building energy use and current commercial practice towards catering this energy expenditure towards the actual, measured needs of building occupants.

It can be seen that the performance gap between designed and in-use building energy rates is a major source of energy waste in the UK, causing a significant contribution towards environmental issues. A wide range of studies have shown the importance of occupants' building use in changing a building's energy demand.

Current building controls systems rarely explicitly measure occupant data, severely limiting their capacity to react appropriately to changing occupant needs and highlight potential places to save energy without negatively impacting occupant experience. In recent years, the Building Energy Management Systems (BEMS) field has seen an increasing interest in the software side of controls: increasing the complexity and intelligence of response to data measured by the systems.

The clear influence of building occupants on dynamic energy demand shows that building energy management, particularly in commercial fields, needs a solid basis for collecting relevant localised occupant data to inform more intelligent controls, sensitive to the limitations of the data collection in an imperfect system. The following chapter presents an examination of how the issues of occupant data collection and its use in building controls have been addressed in existing research work.

3 REVIEW OF RESEARCH ON OCCUPANCY SENSING AND OCCUPANT-CENTRIC BUILDING CONTROLS

3.1 Introduction

While current commercial Building Energy Management systems do not typically collect high-grade occupant data, there is significant research interest in this field. This chapter provides an extensive review on existing research into both the collection of occupant data and its application to building controls.

3.2 Existing Research on the Collection of Occupancy Data

With advances in the scope and availability of various sensing devices in recent years [59], there has been a significant increase in the amount of study into occupancy data in the built environment. Many studies focus on improving sensing technology itself; either through addressing issues with privacy and invasiveness of occupancy sensing or through increasing accuracy of occupancy measurement itself. Other studies explore the use of occupancy data in building services control systems, seeking to improve occupant comfort or decrease energy consumption.

This section covers the methods used by occupancy sensing studies that do not use collected data for any control purpose. Real-time data on building occupancy can also be useful for a range of analytic or user feedback applications, leading to many studies that do not specify how the collected data might be used. The occupant data collected can take several forms, depending on the level of detail needed.

3.2.1 Occupant Presence

The binary parameter of whether or not any occupants are present in a space is one of the simplest forms of occupancy sensing, but is still difficult to achieve 100% accuracy with current technology. Generally, a reasonable accuracy can be achieved by installing motion/PIR sensors, particularly in small spaces with only one occupant [60]. However, PIR sensors require a direct line of sight to the person – it is easy to create blind spots and the sensor must be in plain sight, leading to aesthetic issues [61].

Some research has explored ways to increase accuracy of presence measurement. For example, Hailemariam et al sought to increase the accuracy of sensing presence at individual office cubicles [62]. They used data aggregation from an array of sensors

measuring illuminance, sound level, carbon dioxide (CO_2) concentration, desk power use and PIR motion at each cubicle. It was then assessed which variables added the most value to the presence estimate. Surprisingly, in this study the motion sensor alone performed best, with performance decreasing when more variables were added. This may be due to poor training or noise from the other sensors decreasing overall accuracy.

Multi-sensor solutions have also been tested in domestic settings, such as in the work of Candanedo et al [63], who used localised environmental sensor data to create a model for occupant presence patterns, finding that the highest accuracy was found with 5-minute CO_2 trend data. In some applications however, it is desirable to avoid installing many sensors in an observed space, as this can be seen as intrusive and may involve high costs. A method to estimate presence in a domestic setting was developed using data on electricity use from a home smart meter [64]. This method works on the premise that occupants interact with electrical devices in the home and so occupant presence can be inferred from smart meter data. The use of ambient sensors in a domestic setting is known to have several problems, including motion sensors being falsely triggered by pets/outside events, the slow response time of CO_2 sensors and problem diagnosis issues with a large distributed system of local sensors. The overall accuracy of using smart meter electricity use data was as high as 90.63% in this study. There is a slight issue with false negative estimates, when the occupant is present but not using any electricity. This sensing method is suitable only for small residences and does not provide much useful data for control purposes. It is also possible to exploit the change in air pressure when occupants move between rooms in a residential space in order to sense indoor location to a room level [61]. A transition detection accuracy of 75-80% was achieved using only one sensor in the home.

3.2.2 Number of People

Estimation of the number of people in a space is considerably more complicated than presence alone. Research in the field has explored a wide range of possible methods, each with its own benefits and drawbacks.

The use of a single sensor type to count occupants is attractive due to its lower complexity and installation costs. Several studies have been conducted to determine which sensors give the best approximation of the number of people in a space. A

common option is CO_2 concentration sensors, as the release of CO_2 from respiration is the only major source of CO_2 in most buildings [65]. However, the gas takes some time to build up to high enough concentrations to show elevated occupancy, meaning that response time is slow. Accuracy of counting is also limited by the fact that CO_2 release differs between people and activity levels. Other sensor types used with varying success include heavily-processed temperature readings [66] and volatile organic compounds (VOCs) from use of an office kitchen area [67].

Visual methods use the feed from cameras, which can be specially installed or use existing CCTV equipment, to count the number of people in within the camera's visual field. With strategic placement of cameras, this can allow the number of people in a space to be estimated. Placing cameras only at the entrances and exits of a space allows the number of occupants to be inferred without constantly monitoring the occupants, as in [68]. More detailed techniques can include full context awareness, including the location of tagged objects as well as people [69]. Accuracy of visual systems can depend heavily on lighting conditions, arrangement of furniture and the movement level of occupants. The large amount of data generated by cameras and the need for heavy processing to extract information can make visual systems processing-heavy and slow to run.

The most popular method of people counting in recent research is data mining from several different sensor types. This has the benefit of being able to use non-intrusive, relatively inexpensive sensors and generally achieves higher accuracy than using only a single sensor type. Common sensors included are PIR and CO_2, which are used alongside cameras [68], ventilation actuator signals [70], relative humidity, acoustic and temperature sensors [71]–[74]. In analysis of the information gained by each new sensor type, Lam et al found that the most useful sensors in an open-plan office space were relative humidity, acoustic, CO_2 and temperature sensors [72], while a systematic approach applied by Ekwevugbe et al favoured CO_2 trend, computer use and acoustic levels [74].

Given the current rise of concerns about digital privacy and the collection of data from unknowing participants, combining data from several ambient sensors seems a reasonable compromise from more intrusive detection methods. Data collected through ambient sensor data mining typically does not directly identify the occupant

and holds less sensitive information than directly tracking the precise location of building users.

3.2.3 Real-Time Location Sensing (RTLS)

Location information about building occupants provides a platform for much richer analysis of occupant impacts on building energy use and the possibility of tailoring building control to an individual level. However, this increased detail of collected information causes concerns about privacy and typically requires the specific permission of building occupants to be implemented.

The use of wearable or carried tags to collect location information is not a new concept, but has only inspired research interest in the context of occupant location in recent years. The most common wearable tags use radio-frequency signals to transmit their location to receivers spread around the sensed space [75]–[77]. Radio frequency location systems are divided into sub-sets, depending on the type of radio signal used: active/passive RFID, Wi-Fi, Ultra-Wide Band UWB, Ultra High Frequency UHF etc. When used alone, tagging systems can have some issues with calibration and reflected signals adding noise to the receiver input. It can also be more difficult to detect moving occupants than stationary [78]. The use of many RF receivers around a space [79]–[81] or using RFID in tandem with infrared detectors [82] can significantly improve the accuracy of location, with [79] showing an improved average location accuracy of 93% after adding additional receivers.

Tagging of occupants is only truly applicable to a space where all occupants are known and can be expected to wear a tag – places of work or potentially residences. Ambient sensing can be used to locate occupants in spaces where occupants are not regular visitors, such as public buildings and retail. For example, the interference to wireless LAN signal caused by the human body can be detected and used to infer the location of a person [65], or a dense network of ambient PIR sensors can be used alongside existing CO_2 and humidity data to infer location [83]. Methods such as these are often heavy on computation and may be too slow to run in real-time [83].

As uptake of the smartphone and other smart devices has become more common in recent years, a significant amount of research interest has been put into using personal devices for indoor location [76]. The principle is similar to that of RFID tagging, but avoids the costs and inconvenience of carrying dedicated hardware by using an

occupant's own smart device. There are several possible sources of data that could be used to calculate a user's position:

- Telephone company data/GPS – company data gives a very coarse approximation to location by using a mobile phone's connection to nearby telephone masts, while GPS can offer more detail when enabled on a device. Both methods suffer accuracy issues when used indoors [84].
- Wi-Fi connection – locally distributed Wi-Fi beacons can be used to locate a smart device by requesting connections [84] or the ID of a device's current connected Wi-Fi network can be logged over time [85]. However, some smart device users turn off Wi-Fi when not in use, due to the high power drain caused by leaving it on [84].
- Bluetooth – inquiries by Bluetooth beacons can be a slow process, but it has been shown that speed can be improved by locating devices by which can connect to each other [84].
- Orientation data – the on-board accelerometer and gyroscope found on many smart devices can be used to estimate a 3D path taken by the device, thus allowing the user's location to be inferred if their starting point is known. Systems to improve the accuracy of this method have been developed by combining this data with inter-device connections to provide known points of contact [86] and visual data processing [87].

3.2.4 Activity and Energy Behaviours

It is often an advantage to know the specific activity taking place in an indoor space. This information can be used for user feedback on energy behaviours (e.g. use of appliances, opening/closing windows), analysis of the energy impact of user actions or provide building services control specific to the task's requirements.

As perhaps the least predictable and most varied aspect of occupancy sensing, even when behaviours can be correctly identified, it is important to carefully consider what broader conclusions can be drawn. The energy-related behaviours of a person are influenced by both physical and non-physical parameters [57], meaning that it can be difficult to guess an occupant's intentions and desired outcome even when perfectly recording the physical context in which the behaviours occur. Care must therefore be taken when trying to interpret behavioural data.

A basic approach to learning behavioural context is presented by Bruckner and Velik [32], where a framework is developed for automatically identifying recurring patterns in motion sensor data. It is then assumed that these patterns can be linked to behaviours and occupant intentions by a human observer. Their system builds a Hidden Markov Model (HMM) to build a statistical model of activity level from motion sensors in an office building. Along similar lines, Zhao et al aimed to identify anomalous events in occupancy and their time span, as these events have the greatest impact on the effectiveness of scheduling in building control systems [88]. This approach alone has limited applications due to the need for manual labelling of activities, but presents the first step towards learning behaviour in buildings.

Automated detection of activity types can be achieved by including more detail in the sensed variables. For example, measurements from a range of ambient sensors including motion sensors, sound level and chair pressure pads have been shown to detect office activities with high accuracy [89]. Simpler systems have also been developed using pressure pads in beds and chairs to identify unusual or low activity levels in assisted living spaces in order to identify when an occupant may need medical help [90]. Other approaches include the measurement of when appliances and objects are used [91]–[94], assuming that most human activities involve interaction with some measurable object. A typical problem with this approach is that all expected activities must be predefined and sensors specified for each object interaction. Activities not originally considered will not be detected at all. In a study of multiple domestic buildings [92], it was also found that the accuracy of a trained model using similar data sources varied dramatically from house to house, suggesting some reliance on occupant behaviour matching the predefined activities designed to be caught by the sensors.

Analysis of camera feeds may provide a more generic solution, but requires an extremely high amount of processing and current research cannot produce consistently accurate results [95]. As with any use of processed visual data, there are also issues with privacy that prevent widespread application.

An important point to note is the specificity of occupant behaviour, how personal preferences, position within the room or movement out of an uncomfortable place can change the behaviours occurring. This brings into question the validity of generic

occupancy models that can be applied to any building, and highlights the need to include more comprehensive sensing technology in buildings [57].

3.2.5 Summary and Discussion

It is clear that there is a wide range of sensing technology being tested in current research, both hardware and software. Much of this technology is in its early developmental stages and can be expected to improve in future applications. Table 3-1 summarises the main uses, benefits and drawbacks of the various physical sensors used to detect human occupancy.

When selecting which technologies are most appropriate, a balance between the perception of privacy for the occupant and the accuracy of measurement must be decided. Generally, accuracy increases with the inclusion of more sensors and is highest with the more intrusive options, such as wireless tagging of occupants. Across the studies reviewed, there was a general perception that vision-based systems decrease occupants' sense of privacy, while systems based on passive collection of environmental data are unlikely to be noticed by occupants, and were favoured in studies that valued occupant privacy. Different methods also require different levels of training: combining passive sensors is less intrusive and can give good accuracy, but requires extensive training sets that may not be appropriate for many applications.

The accuracy of occupancy sensing depends heavily on the level of detail measured (e.g. activity sensing gives the highest level of detail but generally has the lowest accuracy). Without specific requirements set by an intended application, it can be difficult to justify what level of detail should be attempted. The following section shows applied examples of occupancy sensing, with further justification of the balance between accuracy and detail.

Table 3-1 - A summary of sensors used in occupancy detection and their uses

Technology	Level of Detection Possible	Strengths	Weaknesses	Ideal Applications
Passive Infrared (PIR)/ Motion sensors [32], [62], [65], [67], [68], [70]–[74], [83], [85], [92], [96]	Occupant presence (alone), Number of people (combined)	- Relatively low cost - Readily available - Less intrusive	- No counting capability - False negatives when occupants are still - Require direct line of sight	- Single-person offices - Individual cubicles
CO_2 [62], [63], [65], [67], [68], [70]–[74], [83]	Occupant presence, Number of people	- Readily available - Non-intrusive	- Slow response time - Affected by ventilation	- Smaller volume spaces - Known activity level
Volatile Organic Compounds VOC [67]	Occupant presence, Number of people	- Can detect activity-specific person count - Non-intrusive	- Very specific application	- Kitchen areas
Smart meter data mining [64]	Occupant presence	- Uses existing infrastructure - Non-intrusive	- False negatives when occupants are not using electricity	- Residential
Illuminance [62], [73]	Occupant presence	- Relatively low cost - Readily available - Non-intrusive	- Must be combined with other sensors	- Ambient sensor combination
Acoustic [62], [71]–[74]	Occupant presence, Number of people	- Relatively low cost - Readily available - Less intrusive	- Must be combined with other sensors	- Ambient sensor combination
Appliance/lighting use [62], [67], [74], [85], [92], [94]	Occupant presence, Number of people, Activity	- Non-intrusive	- Misses occupants/ activities not using electricity	- Ambient sensor combination - Workplace activity sensing
Temperature [63], [66], [71]–[73]	Occupant presence, Number of people, Location (combined)	- Relatively low cost - Readily available - Non-intrusive	- Must be combined with other sensors	- Ambient sensor combination
Door Open/Close Status [85], [96]	Occupant presence, Number of people, Activity	- Non-intrusive	- Must be combined with other sensors	- Ambient sensor combination

Door Counter [88]	Number of people	-	Non-intrusive	-	Can be skewed by multiple people at once	- -	Public spaces Workplaces
Humidity [63], [73], [83]	Occupant presence, Number of people, Location (combined)	- - -	Low cost Readily available Non-intrusive	-	Must be combined with other sensors	-	Ambient sensor combination
Cameras [68], [69], [87], [95]	Occupant presence, Number of people, Location, Activity	-	High level detail possible	- -	Privacy concerns Heavy processing required	-	Workplaces, public places with existing CCTV
Radio Frequency tags [78], [79], [82]	Location	-	High level detail possible	- -	Privacy concerns Hardware must be carried	-	Workplaces
Pressure Pads [85], [89], [90], [92]	Location, Activity	-	Can monitor specific location of interest	-	Privacy concerns	-	Assisted living, domestic, office
HVAC Actuation	Occupant presence, Number of people	-	Helps to account for ventilation effects	-	Relationship to occupancy can be indirect	- -	Workplaces Existing HVAC systems
Air pressure change [61]	Occupant presence, Location	- -	Non-intrusive Can sense movement between spaces	-	Relationship to occupancy can be indirect	- -	Low occupancy spaces Residential
Smart device tracking [84]–[87]	Location	-	High level detail possible	- - -	Privacy concerns Hardware must be carried Assumes all occupants have a device	- -	Workplaces Residential

3.3 Application of Occupancy Data

When occupancy data is applied for a specific purpose, it is easier to define the requirements from sensing equipment. Development of building control algorithms using real building data have been both driven by advances in sensor technology and a driver for new sensor types [97]. Studies that use applied occupancy data in the control of building services can be categorised into four levels [98]:

- Control using real-time occupancy data

- Accounting for the preferences of individual occupants
- Control informed by prediction of future occupancy levels
- Control catered to individual activities

3.3.1 Control in Real-Time Response to Occupants

The use of real-time response to building occupancy is well established in commercial lighting control, with widespread use of PIR sensors to switch on/off lighting. Some research has been dedicated to improving the efficiency and user convenience of such lighting systems. For example, Garg and Bansal [99] showed that altering the sensor's time delay through the day according to expected activity level can produce a small energy saving and reduce unwanted lighting switch-off, while Labeodan et al showed that a relatively simple system of motion and chair sensors could save energy on localised office lighting [100]. More significant energy saving was shown by Xu et al [101], where real-time complex event processing was used to control localised lighting in a system designed to understand how office space was used and be robust to changes in equipment or sensor layout.

Along similar lines to lighting control, occupancy data can inform control of electrical power supply to appliances; saving energy by automatically switching off power to appliances when not in use. Domestic solutions have been tested [102], but commercial applications in multi-occupancy spaces remain more complex. Some studies have focussed on local motion sensing to detect presence close to appliances [103], while others have used Bluetooth tagging systems to detect when the owner is approaching equipment, allowing the equipment to reboot in time to be used [104]. However, it was concluded that more context was required for reliable appliance control.

As detailed in section 2.4, the control of building HVAC systems through BEMS is a rapidly developing field. As such, there is much research interest in improving the response of HVAC to occupants in real-time. Operations to reduce energy use from real-time data include [78]:

- Maintaining temperatures further from the set point in unoccupied areas,
- Maintaining lower ventilation rates in unoccupied areas,
- Supplying airflow based on occupancy,
- Adjusting outside air volume based on occupancy,

- Responding to dynamic heat loads on a timely manner,
- Operating HVAC systems based on occupant preferences,
- Learning energy consumption patterns,
- Increasing the flexibility of control.

Most research in this field focuses on commercial applications, making use of the prevalence of existing building sensing and automation equipment in the commercial sector. However, some research shows that in residential applications there is potential for significant energy saving from sensing occupant presence in real-time, rather than using a typical fixed HVAC schedule [105]. Meyer and Rakotonirainy provided an overview of context-aware home projects and their differences to commercial applications [106]. In particular it was noted that commercial and home applications of automation will have different sets of priorities, with homes catered to comfort while commercial systems focus on productivity.

The focus on improving productivity while reducing energy is particularly true for office environments: an area with significant existing research. The benefits of real-time occupancy-based HVAC control vary significantly with the location and building utilisation: up to a 56% saving using simple CO_2-based estimation of number of people in a busy Hong Kong office [107]. Yang et al developed a model to detect presence and the number of people in office space, using ambient sensors. This sensing was used to simulate the potential energy saving from demand-based HVAC control [108] and build models of occupancy for simulation [109]. Agarwal et al have conducted a range of studies on office applications, simulating an energy saving potential between 10-15% relative to a conventional HVAC control system in single-person offices [110]. Comparison of system complexities suggest that single-occupancy offices gain significant benefit from simple occupancy or CO2 responsive HVAC, with a smaller marginal benefit of more complex strategies [111]. It is demonstrated that the combination of PIR and door open/close sensors can yield 96% accuracy in single-occupancy office spaces [112], providing greater accuracy than Wi-Fi smartphone tracking [113], but only suitable for single-person spaces. When applied to a real building with varying space types, energy saving from their Wi-Fi based control system was higher, at 17.8%. In a similar vein, Zeiler et al [114] used wireless tags to locate occupants in a multi-use office space, demonstrating energy saving with localised heating control.

Occupant-driven HVAC can be particularly effective when combined with other sustainability technologies. For example, Rosiek and Batlles used PIR and CO_2 sensors to inform the energy use and storage from renewably driven air conditioning [115]. In a building with highly variable use levels, energy saving of 42% was demonstrated by accounting for occupant presence.

3.3.2 Control to Individual Occupant Preference

Some control applications benefit from comfort input from individual users, catering building conditioning to each person's own preferences. This typically requires one of three ways to estimate the comfort level of building occupants:

- Voluntary information provided by the user,
- Estimation of comfort from empirical models and environment data,
- Attempts to measure user comfort automatically.

Once collected, preferences can be applied to a range of building controls. Personalised lighting levels at a desk surface can be achieved by controlling lamp output using occupant preference data, natural light levels and a physical relationship between the lamp settings and work surface illuminance [116].

More commonly, occupant preferences are collected and used for control of thermal systems in a building. Thermal comfort is highly subjective and depends on a wide range of variables [117]. In theory, better knowledge of a person's comfort level should allow much more responsive HVAC systems, greater overall comfort and improved energy efficiency. The use of a calculated 'predicted mean vote' comfort level among occupants was demonstrated by Zhao et al [118], Kolokotsa et al [119], and Gao and Keshav [120]. The latter developed a system to control heating supplied to individual desks in an office space. Here, occupant preference was estimated by calculating a modified version of 'predicted mean vote', an empirically-based value for expected comfort level. This calculation was informed by local air temperature, humidity, air speed, radiant temperature and clothing level. Equipment for the system was costly and required much calibration.

A more specific way to implement user preference into control is to record and maintain a set of preferences for a specific individual while tracking their location in the building. This can be achieved by: surveying occupants at their single-occupancy

desks [121]; using personal smart devices to locate users and perform periodic comfort surveys [122]; using a range of ambient sensors to locate users while recording manual control adjustments [123]; or RFID tagging and recording manual adjustments to building controls made by each user [124], with the inclusion of PIR for greater accuracy of location sensing [125]–[127]. In multi-occupant spaces, distinction can be made between groups with tracked locations/preferences and unknown occupants detected by ambient sensors [128]. It has also been demonstrated that occupant skin temperature observed with an IR camera can be used as a basis to determine comfort levels and adjust local conditioning [129].

It should be noted that some of the above applications are highly specific to the type of space in which they are implemented: the iDorm project [123] used a range of sensors that are only applicable to learn personal preferences in a residential context with a single main occupant. The learning of rules is designed to occur based on the exact context of the occupant at the time (sleeping, working etc.), requiring a system that is more detailed, but less widely applicable to other buildings and spaces. By contrast, the work of Moreno-Cano et al [125]–[127] is designed to be a generic platform that can learn the requirements of each user in any type of space. While less detailed than the iDorm project, this application proved to be effective and yielded an average energy saving of 20% during testing in a multi-use building.

3.3.3 Control to Individual Behaviours/Activity Types

Although much study has been conducted on the detection of occupant activities, this information has not often been implemented into control systems. This is partially due to continuing problems with detection accuracy (see Section 3.2.4) and the lack of hardware required to control building systems finely enough to cater to individuals' activities.

Some initial experiments into activity-based control include optimisation of localised lighting [130], in which cameras are used to define whether occupants are working, moving, sleeping etc. and choose appropriate lighting configurations for each activity. While energy saving was demonstrated, the more energy-efficient control options were perceived as less useful lighting by occupants, suggesting that the system was not successful enough for application.

A simpler approach was demonstrated in [131], where individual office desks were conditioned only when the occupant was present. Control inputs for lighting, appliance power supply etc. were adjusted when it was sensed that the user was working on their computer/working on paper. This strategy is only possible in a space with office desks that are sufficiently isolated from one another. The issue of detecting multiple occupant activities in one zone was also encountered in a domestic study [132], where motion and in-room location were manually mapped to predefined activity types. When the start of an activity was identified, the heating system was adjusted in response. If continuation of the activity was then confirmed by observation, the heating strategy continued. However, the system was not built to respond to conflicting activities within a single zone.

3.3.4 Control through Occupancy Prediction

While the studies in Section 3.3.2 attempt to estimate the control action needed given a particular context, research in this section tries to predict the context of a space in the near future in order to pre-emptively condition the space to acceptable levels. In theory, this allows areas to drift further away from comfort conditions while not occupied, saving energy that would have been used keeping a space to 'standby' comfort levels in case it becomes occupied in the near future. This control method offers a step beyond simple reaction to an occupant's presence and generally requires a more complex control system.

Studies into occupancy prediction can be categorised by exactly what parameters are being predicted and the intended horizon of prediction. For applications that require a fast response, the prediction horizon is typically low, in the order of minutes/hours [133] or simply concerned with predicting the next action in a sequence. This form of prediction is useful for the control of power supply to household appliances, lighting etc. as in the Adaptive House [134], [135] and MavHome [102], [136] projects, where the environment around a user records typical sequences of actions during use of the home and then attempts to automate more common sequences.

For application to systems with a slower response time, such as HVAC control, predictions of occupancy must run over a longer time horizon. This requires the construction of a model of occupant behaviour from extended observation of a space. It has been shown that the predictability of occupancy is dependent on how the space

is used [133]. However, there has been some degree of success in this field. For example, Mamidi et al [137], [138] and Howard et al [139] developed prediction from the current state, up to 90 minutes in the future. Between the two studies, several different prediction methods were tested, trained with ambient sensor data from two educational buildings. Accuracy of prediction using real building data from a large educational space was high, ~90% for Mamidi et al. It was found that more variation in short-term occupancy gave lower accuracy of prediction. Other applications have sought to predict the duration of the currently observed binary occupant presence status by comparing to the most similar situations in previous observations using a k-Nearest-Neighbours algorithm [140]. Identified issues with the systems above include the need for a large amount of training data.

Dong et al built upon their previous work on counting the number of people in a commercial space [71] to develop a system to predict the duration of the current occupancy state from historic event sequences inferred from ambient sensor data [141]. An advantage of this approach is that it requires no training period, simply starting learning when it is implemented. The predictive control was then refined to include office calendar data, weather data and a physical model of the building for fully optimised proactive control [142], [143]. This proved to save 18-30% energy in implementation on a real building. A relatively simple system of motion sensors was used with a stochastic algorithm to predict occupancy duration and highlight absent days for lower heating setpoints in the work of Gunay et al, with simulated heating savings of 10-15% [144]. Another approach to online learning was proposed by Dobbs and Hencey [145], who used sensor input to gradually build a predictive model of occupant presence, weighting newer data to account for changes in building use over time. Aswani et al [146], [147] produced a thermal model that did not require manual training, as the thermal impact of occupancy was automatically identified by the model. This model predictive approach showed 30-70% energy saving over simple temperature control.

Domestic HVAC is also the subject of predictive control research [96], [148], [149]. In [148], daily profiles of use and comfort preferences were created from observation in a test home fitted with several sensor types. A daily profile from an observed set was selected based on home use the previous day and corrected if observation no longer matched the selected profile. In [149], the cell network and Wi-Fi connections

of personal mobile devices were used to measure the room-level location of domestic occupants, which then was used in a Markov predictive model to estimate transitions between rooms. It was found that the system was accurate while occupants behaved within typical routines, but was not able to handle unusual behavioural patterns.

It is also possible to predict occupancy over several timescales in order to better account for unusual behaviours. The work of Erickson et al uses short prediction windows to ensure infrequently occupied spaces are conditioned and longer windows to ensure that frequent but short visit spaces are conditioned. Erickson developed predictive HVAC control strategies over several years, with testing of several different methods for prediction of occupancy and control algorithms based on such prediction. Early work [150], [151] focussed on comparing two prediction models informed by a network of cameras: a multivariate Gaussian distribution model (MVGM) was used to generate a coarse prediction of when a room is empty or occupied, while an agent-based model (ABM) was used to simulate the route taken by individual building users. It was found that ABM produced good simulation of building use for design purposes, but was not suitable for predicting future room usage. MVGM was determined to be better for real-time prediction, as it is informed by current occupancy states, but poorly represented certain behaviours, such as underrepresenting rooms that are not often occupied.

In response to the limitations identified above, a system based on a Markov Chain model was developed for predictive HVAC control [152], [153]. After training with ground truth data to build a probability matrix of transitions between states, the current occupancy state is used to predict occupancy per room over the next several time steps. These steps are then used to inform whether the space should be conditioned to comfort or setback temperatures. In simulation, this system is shown to reduce energy use by up to 42% relative to fixed HVAC schedules. The energy saving and comfort provided by occupancy prediction is noted to be higher than a purely reactive strategy. It is also shown that knowledge of the number of occupants in the space provided greater energy saving for ventilation than binary presence. The prediction method was then tested in a real-life application [154]. An estimated annual energy saving of 30.0% from trial results suggests that the simulation work was optimistic, but essentially sound.

The use of occupancy data to predict internal gains over the long term (up to 60 hours) for the purpose of optimising HVAC suggests that energy can be saved without making changes to set points when a room is vacant [155].

While many studies show the benefits of occupancy prediction in appropriate applications, there is some debate about the value of including complex predictive control over simple reaction. Oldewurtel et al tested the benefit of prediction over real-time reactive control by simulation of a single-occupancy office environment [156] [157]. Prediction was used to determine the likelihood of occupancy on a given day, offsetting the HVAC load when the room is vacant all day. It was concluded that predictive control did not provide a significant benefit above reactive control. Along similar lines, Goyal et al tested several HVAC control methods, including fixed schedules, reaction and prediction through simulation [158] [159] and experimentation [160] for a small office space. Both reaction and prediction improved response relative to the baseline, but prediction did not show significant improvement over reaction. Inclusion of a building thermal model reduced oscillation, but was sensitive to uncertainties in the input data for the plant model and occupancy [161]. This suggests that more robust occupancy measurement technology would further strengthen the case for predictive control.

It should be noted that in the above studies, the tested spaces are generally small and are not expected to have many occupants. Other studies have noted that spaces such as this, which can be conditioned quickly to comfort conditions, are not ideal candidates for predictive control [120] [162]. Simulation of different room sizes and constructions shows that prediction can be more useful in larger spaces with higher thermal mass [161], but this effect is not fully explored in this paper. In spaces with slower response times, it has been shown that effective preconditioning of the space can yield significant energy savings [163]. Indeed, in a review of relative savings from various control strategies, those with occupancy prediction showed the greatest percentage energy saving [59].

3.3.5 Discussion

Table 3-2 to Table 3-5 show a summary of the occupancy-based controls discussed in this section. The vast majority of studies show that accounting for occupancy in the control of building systems can significantly reduce the energy used in a building.

Occupancy data is used in the control of appliance power supply, electrical lighting, automated blind placement, heating, ventilation and air conditioning. The amount of energy saved varies dramatically between applications and control intentions.

Due to the wide range of building types and control methods applied, it is difficult to truly compare the energy saving made by different studies. In research that directly compares different strategies in the same building, it appears that the greatest overall energy savings can be achieved with controls that combine reactive and predictive approaches to optimise conditioning of a space. However, there is debate on whether the benefits of predictive control justify the significant increase in complexity and computing power required over purely reactive control. In general, the value of predicting behaviours over real-time response depends on the application – most prediction models are applied to HVAC operation, as this typically has the slowest response of any building system and so can benefit the most from early warning of demand changes. This is discussed in more detail in Section 3.3.4.

Controlling building systems to individual preference is shown to create issues with conflicts in multi-occupancy rooms. Energy saving varies on the preference of the occupant, but comfort levels should be held as high as possible for all occupants. Further research may be required on how best to balance conflicting comfort standards in a multi-occupancy space.

An attribute common to most studies in the above section is that the control is often tested through simulation, rather than deployment in a real building. Given that there are many identified issues with current simulation technology, there would be significant worth in testing systems by implementation in real-life buildings where possible.

It should also be noted that the comfort of occupants should be accounted for, as well as direct energy saving. With the rise in sophisticated automated controls, it is worth considering that occupant comfort is highly subjective and can often be related to the illusion of control: placebo-like effects have been demonstrated with dummy controls [57]. This highlights the need to check that occupant satisfaction is not negatively affected by energy saving measures.

Table 3-2 - Summary of real-time occupancy-based control studies

Authors	Control Type	Occupancy Data Used	Techniques	Building Type	Energy Saving Demonstrated
Garg & Bansal [99]	Lighting	Occupant presence - PIR		Office	25%
Labeodan et al [100]	Lighting	Local Presence - PIR - Chair Status	Rule-based control	Office	24%
Xu et al [101]	Lighting	Occupant presence - PIR - Door Status - Light Barrier	Complex event processing	Office	≤ 34%
Park et al [103]	Appliances	Occupant presence - PIR		Office	≤ 21%
Harris & Cahill [104]	Appliances	Real-time location - Smart device tracking		Office	Not reported
Batra et al [105]	Domestic HVAC	PIR/Door status		Domestic	Not reported
Sun et al [107]	HVAC	Number of people - CO2		Office	≤ 56%
Agarwal et al [110]	HVAC	Occupant presence - PIR - Door status		Single & multi occupant offices	10-15%
Agarwal et al [113]	HVAC	- Smart device tracking		Single & multi occupant offices	17.8%
Zeiler et al [114]	HVAC	Real-time location - Wireless tagging	Zone allocation	Multi occupant offices	30-45%
Rosiek and Batlles [115]	HVAC	PIR/CO2		Lab/office/ Meeting room	42%
Gruber et al [111]	HVAC	Number of occupied offices/CO2	Simple totals vs multi-objective optimisation Energy Simulation	Single occ office	6-39%
Yang et al [108]	HVAC	Number of people - Ambient sensors	Decision Trees for detection Energy simulation	Office/ Classroom	18-20%

Table 3-3 - Summary of preference-based control studies

Authors	Control Type	Occupancy Data Used	Techniques	Building Type	Energy Saving Demonstrated
Singhvi et al [116]	Lighting	- Illuminance - Appliance/Lighting use - RF Tags			Not reported
Zhao et al [118]	HVAC	- Temperature - Surveying	Predicted Mean Vote comfort model Measured MV	Office	44%* *Simulated result
Kolokotsa et al [119]	HVAC, Lighting, blinds, window opening	Individual presence and comfort - Temperature - Humidity - Air Speed - CO2 - Illuminance	Predicted Mean Vote comfort model Fuzzy controller	Single-occupancy office	Not reported
Gao & Keshav [120]	HVAC	Individual presence and comfort - Temperature - Humidity - Air Speed - Cameras	Predicted Personal Vote comfort model	Office	Not reported
Jazizadeh et al [121]	HVAC	Individual presence and comfort - Temperature - Comfort surveying	Fuzzy Logic for profile construction	Single-occupancy office	26-39%
Yong et al [122]	HVAC	- CO2/Temperature/ Humidity - Illuminance - Acoustic - Air Speed - Smart device comfort surveying	Individual agents	Office	Not reported
Hagras et al [123]	Domestic HVAC, lighting, blinds	- PIR - Illuminance - Appliance use - Temperature - Pressure pads - Window opening	Fuzzy logic, individual agents	Domestic	Not reported
Chen et al [124]	Lighting, HVAC	Real-time location - Appliance use - RF tags		Office	Not reported
Moreno-Cano et al [125]–[127]	Lighting, HVAC, switches, blinds	Real-time location - PIR - Appliance use - RF tags	Radial Basis Functions Neural Network	Office	20%
Yeh et al [128]	Lighting, HVAC, switches	Real-time location - RF tags - Light sensor - Acoustic - Temperature		Office	16.5-46.9%
Vissers and Zeiler [129]	HVAC	Individual comfort - IR cameras - Ambient Temperature		Office	<17%

Table 3-4 - Summary of activity-based control studies

Authors	Control Type	Occupancy Data Used	Techniques	Building Type	Energy Saving Demonstrated
Xu et al [101]	Lighting	Occupant presence - PIR - Door Status - Light Barrier	Complex event processing	Office	≤ 34%
Lee et al [130]	Lighting	Activity type - Cameras	Conditional random field model	Generic	Not explicitly reported
Milenkovic & Amft [131]	Lighting, power supply	Activity type - PIR - Appliance use at desk	Finite state machines, Probabilistic layered hidden Markov models	Office	21.9%
Pallotta et al [132]	Heating	Activity type - PIR - IR camera - Accelerometer		Domestic	Not reported

Table 3-5 - Summary of occupancy prediction-based control studies

Authors	Control Type	Occupancy Data Used	Techniques	Building Type	Energy Saving Demonstrated
Harle & Hopper [133]	Lighting, Appliances	Real-time location - Ultrasonic location	Ingress zones for lighting/appliances.	Office	50% Lighting only
Mozer [134], [135]	Lighting	Activity type - Ambient sensors - Appliance use	Neural network predictor	Domestic	Not reported
Cook et al [102], [136]	Appliances, lighting	Activity type - Ambient sensors - Appliance use	Modified text compression algorithm	Domestic	Not reported
[137], [138] Mamidi et al	HVAC	Occupant presence Number of people - Ambient sensors	Multilayer perceptron and logistic regression classifier	Lab/Office	Not reported
Howard & Hoff [139]	Not reported	Occupant presence - PIR network	Bayesian combined forecasting	Lab/Office	Not reported
Peng et al [140]	HVAC	Occupant Presence - Motion sensors	k-Nearest Neighbours	Single & Multi Office	20.3%

Dong et al [141]–[143]	HVAC	Number of people - Ambient sensors	Online model building, machine learning, MPC	Office	18-30%
Gunay et al [144]	HVAC	Occupant presence - PIR network	Online stochastic model	Office	10-15%
Dobbs and Hencey [145]	HVAC	Occupant presence	Online model building, Markov chain, MPC		8%
Aswani et al [146], [147]	HVAC	Number of people - Temperature	MPC	Computer lab	30-70%
Barbato et al [148]	Domestic HVAC	Occupant presence - PIR network		Domestic	≤ 28%
Lee et al [149]	Domestic HVAC	Occupant Presence, Room-Level Location	Markov Chain	Domestic	Not reported
Erickson et al [150]–[154]	HVAC	Occupant presence, Number of people - Camera network	Agent based modelling, Multivariate Gaussian model, Markov chain model	Lab/Office / Meeting room	≤ 42%
Zhang et al [155]	HVAC	Number of people as part of internal gains	MPC variations	Single-occupant offices	Not reported
Oldewurtel et al [156] [157]	HVAC	Occupant presence - PIR	MPC	Single-occupant offices	≤ 34%*
Goyal et al [158]–[161]	HVAC	Occupant presence Number of people	MPC	Single-occupant offices/ Meeting room	≤ 56-61%* *Simulated result of ideal prediction

3.3.6 Simulation and Testing

Occupancy data can also be useful for improving the accuracy of simulation, energy prediction and testing of design hypotheses. The energy demands and overall consumption of a building can be significantly affected by occupancy. It is therefore essential that simulation and testing use realistic occupancy patterns for accurate

results – a need that is not fulfilled by the static occupancy schedules typically input to current simulation tools.

There have been several studies that attempt to use large datasets of how a building in used to derive a model to generate any amount of typical occupancy data. In particular, domestic Time Use Survey (TUS) data has been used to build stochastic models that can generate daily occupancy schedules based on household size [164]–[170]. Most of these studies find their basis in the Richardson model [168]–[170], available online. In this model, UK TUS data is used for energy prediction as an aggregate for the whole house, based on estimated activities. The load for HVAC is not included. Additions include consideration of duration of activity [165] [167] and detailed estimation of HVAC load [166]. Meidani and Ghanem [171], [172] attempted to derive a similar model from a much smaller input data set, accounting for the higher uncertainty present when less data is available. It is noted that transition matrices based on observed data will always have some error and variability: the best that can be done is to minimise the error, rather than eliminate it.

In commercial settings, the lack of large-scale TUS data means that models have been developed through smaller scale measurement of more specific processes. For example, office-based small power energy use was estimated by two different models developed by Menezes et al [173]. The first model used random sampling from detailed measurement of the use of office equipment, while the second used coarse data on the number of appliances and office schedules to estimate use profiles. Both approaches had some inaccuracies, including underestimation of peaks, which may be detrimental to design use and demonstrate the need for more data collection in the field. Models have also been developed for interaction with natural ventilation [174] and the presence of occupants at office cubicles [175]. Here, the model uses data on the frequency, duration and starting time of absence periods to create daily presence profiles that fall within typical bounds of building use.

Lee and Malkawi made use of agent-based modelling to estimate window use behaviours for simulation of commercial buildings [176]. The comfort of occupants was calculated and used to predict when windows are opened/closed over time. Use of this model in simulation resulted in significantly different internal temperatures predicted relative to standard simulation methods. This agent-based approach was

then extended to include other behaviours for the simplified case of a single-person office [177]. In order to address uncertainties in real-world application, it was assumed that human behaviour is logical according to the perceptions and expectations of the individual.

Liao et al developed an agent-based model to simulate the location of occupants over time [178], [179]. This allows the number of people in each arbitrary zone to be predicted at any time, given the initial conditions. The model developed in this study is easily scalable to an arbitrary number of zones and agents. However, it is not suitable for real-time estimation in this form. Surveying, scheduled activities and access rights of building occupants are used to build profiles for each agent. Agents are assigned a primary zone, in which they will spend most of their time. The model is calibrated by comparing the results to measurements for whole-building mean occupancy etc. Testing showed that mean occupancy was overestimated and daily entry/leaving times were not accurate, but general accuracy to real building use was acceptable.

Most of the above studies cover Monte Carlo simulation – running a deterministic model many times with stochastic inputs. This creates a distribution of the output variable over many runs of the model. An alternative method is to model energy use over time in two parts, such as in the work of Brohus et al [180]: here building loads were modelled as a function of the mean value over time, with an addition of fluctuating noise to represent randomness from occupant behaviours, weather fluctuations etc.

Another study [181] utilises in-use measurements to predict energy consumption of buildings. On-line models do not require the high volumes of training data that is needed for accuracy in static models. Regression models are suited to estimating average energy use, not specific use in small time steps. Accumulative training (constantly revising model with input data) can allow both local and global trends to be found. However, over time the volume of data may slow the running of the system. Newer trends also have a smaller impact relative to the rest of the data. Sliding window training does not have problems with aggregating large volumes of data. Older data is discarded over time. This could lose longer-term trends. It is also difficult to choose the optimum window size. The MATLAB ANN toolbox is used to

develop this paper's model, with several pre-set algorithms for training tested. Despite stating the need for on-line training, the adaptive models did not significantly outperform the static models.

3.3.7 Discussion

The differences of approach between control-oriented and design-oriented applications of occupancy data reflect the different requirements of each process. Control must be quick to process and typically works on a short horizon – between minutes to a day.

Design data is typically used for long-term simulations, so is not constrained by processing time. It needs to generate a large dataset that accurately represents the mean and variation of real data, without the need for accuracy when comparing particular days. This means that it is much easier to extrapolate design data from a historical dataset than to predict specific behaviours that will happen on a particular day, as is needed for control. This highlights the need for continual real-time updating of context for control models, so that both reactive and predictive actions are accurate and provide useful results for energy saving and comfort.

3.4 Conclusions

From the wide range of approaches into the collection of occupant data, it can be seen that the measurement of building occupants is a highly complex problem that requires a clearly defined aim in the type and level of data collected, as well as a necessary trade-off between the level of detail measured and the perceived intrusion into occupants' privacy. 'Occupant data' could encompass a wide range of actual data types, from simple binary presence to highly computationally intensive systems to infer specific occupants' activities.

The review of studies applying occupancy data to building control has highlighted areas of interest within the field. The existing body of research has tested the application of occupant-centred building control across a range of building services and to varying levels. Applications were split broadly into four categories: systems that respond in real time to occupants, those that collect and utilise data on individual occupant comfort preferences, controls catered to specific occupant activity and proactive control based on prediction of future events. In particular, the study of

predictive energy management and control was found to have significant potential for energy saving, particularly from systems with a slow response time, where real-time response alone may result in energy waste or uncomfortable conditions as the system catches up to the demands of the current occupancy state. In research that compared different strategies in the same building, the greatest overall energy saving was achieved with controls that combined reactive and predictive approaches to optimise conditioning of a space, although some authors questioned the added value of prediction versus its increased computational requirements.

It is proposed that a combined responsive-predictive strategy shows the greatest potential for improving the efficient control of building systems towards actual occupant energy requirements. However, it has been demonstrated that the field of occupancy prediction and its application to control is a problem that has not been exhaustively solved. The focus of this research work was therefore directed towards the development of a system that could supply useful predicted occupancy data for building control in multi-occupant spaces, with sensitivity to the fact that future events will never be predicted with perfect accuracy. As a basis for this development, data sources needed to be selected for the feasible collection of occupant data and analysis to target the type of situations where occupant-centric energy saving measures could provide the most benefit.

4 POST-OCCUPANCY CASE STUDIES OF OCCUPANCY AND ENERGY

4.1 Introduction

As an introductory work for this study, a series of case studies across sectors were used to assess the current state of applied occupancy data collection and to make initial assessments of the existing relationship between occupancy rates/behaviours and energy usage.

Data sets of some measure of occupancy and energy were acquired for two case studies: a domestic setting based in Nottingham, UK and a large office building based in Worksop, UK. Some qualitative occupancy-related assessment was also conducted for a school building in Dagenham, UK.

4.2 Green Street Domestic Dataset

4.2.1 Building layouts, uses and data types available

The Green Street housing project is based in the Meadows area of Nottingham, UK, situated as shown in Figure 4-1. The houses were developed in three 'phases', with slight variations on building design between phases as shown in Appendix 10.1.

Historic energy use and basic occupancy data was available from eight houses across all three phases. The investigation of the Green Street data focussed on several key questions:

- How well can occupancy be quantified using motion sensor and CO_2 data with no ground truth available to verify results?
- Does energy use relate to occupancy? Which types of energy have the closest correlation?
- If types of energy use correlate directly with occupancy, can they be used to add reliability to simple motion sensor data?
- Does the behaviour of occupants in different houses cause quantifiable energy waste relative to other houses?

Figure 4-1 - Green Street Project, Meadows Area, Nottingham, UK

The data types available and sensor locations for the eight houses are summarised in Table 4-1. Data was collected via wireless sensor nodes, detected by a centralised hub and sent to a central storage point. This allowed web access to the data in CSV format and was maintained by a contracted company during the study. However, the nature of the centralised wireless detection caused some issues with data fidelity, discussed below. The layout of the house sensor equipment differed between houses A, B, C, D from Phase 1 and houses E, F, G, H from Phases 2 and 3. Diagrams of both house types can be found in Appendix 10.1.

Table 4-1 - Occupancy and Energy Data available from the Green Street Project

Information Type	Data Name	Description	Measured Units	Houses Available
Occupancy	Footfall – Downstairs	Number of times the downstairs hallway PIR motion sensor was triggered	Count (whole numbers)	All
	Footfall – Upstairs	Number of times the downstairs hallway PIR motion sensor was triggered	Count (whole numbers)	All
	CO_2	CO_2 concentration in living room	ppm (parts per million)	C, G
Electrical Energy	Main Electric Import	Total electricity imported to house on top of any generated electricity	kWh	All
	Main Electric Export	Total electricity exported from house from generated electricity	kWh	All
	Extraction System/Heat Recovery	Electrical energy used by the house ventilation system	kWh	All
	Hob/Cooker	Electrical energy used by the cooker	kWh	All
	Sockets Downstairs	Phase 1 - 1st Floor and Kitchen, Phase 2&3 – Ground Floor	kWh	All
	Sockets Upstairs	Phase 1 – 2nd Floor, Phase 2&3 – 1st and 2nd Floor	kWh	All
Heating Energy (Gas boiler)	Heating	Energy used in heating the house	m^3 for A, C, D and kWh for B, E-H	All
	Main Gas	Gas used by the boiler for heating and hot water	m^3	All
Water Use	Hot Water	Hot water use in the house	m^3 for A, C, D and kWh for B, E-H	All
	Mains Water	Total water use in the house	m^3	All

4.2.2 Data Fidelity

A visual comparison of data from the motion sensors for each of the houses – shown in Figure 4-2 and Figure 4-3 – indicated that there were common periods of sensor dropout, indicated with the shaded sections of the figures. Features common to these sensor dropout periods included:

- Timing of dropout is common to buildings in Phase 1, different timing for Phase 2 buildings.
- Dropout was observed across most sensor types at once.
- Dropout was immediately followed by an unfeasibly high peak. Further examination showed that the peak was approximately equal to the sum of expected readings over the dropout period.
- Dropouts have consistent zero readings.

Phase 1 houses showed the largest dropout period, lasting from 14th August – 3rd October 2013 and affecting all sensor types from the Phase 1 houses. Phase 2 buildings showed different periods of common sensor dropout – most notably the period 14th – 23rd February 2014. House G additionally had no PIR readings from February 2014 onwards, indicating a likely sensor failure within this building.

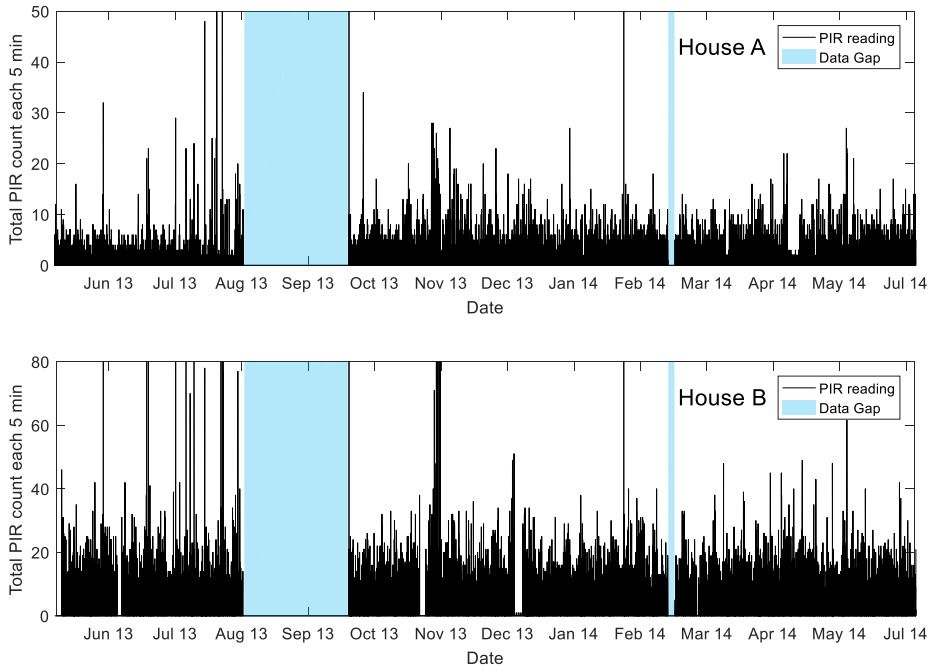


Figure 4-2 - Common Data Dropout Periods in PIR data for Phase 1 buildings A and B

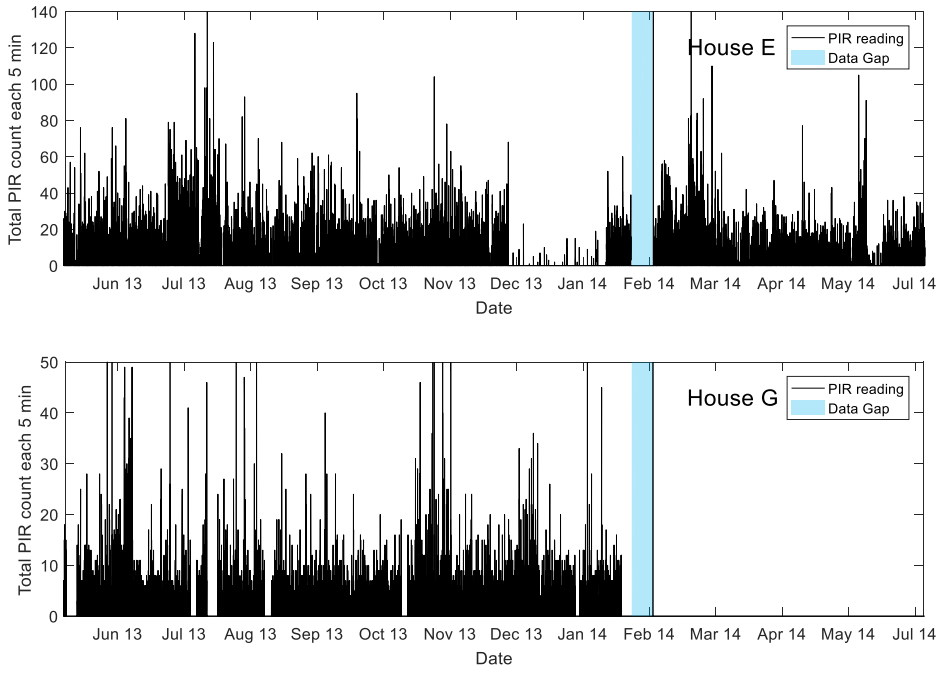


Figure 4-3 - Common Data Dropout Periods in PIR data for Phase 2 buildings E and G

Given the consistent nature of the dropout periods, it was possible to write a script to flag any instance of sensor dropout, allowing for identification of shorter dropouts than could be found with manual inspection of the data. This algorithm is described in pseudocode below.

Algorithm:

Start indicator set to zero

For each 5 minute timeslot of data
    If all sensors read zero and start indicator is zero
        Note the timeslot at which zeros start in start indicator
    Else if start indicator is not zero and extractor data peaks in this or next timeslot
        Note the timeslot at which the peak occurs
        Add start indicator to list of dropout start points
    Else if start indicator is not zero and there is no peak
        Return start indicator to zero
    End
Next timeslot

The algorithm was run on data from house C in order to verify its success rate. A total of 183 dropouts were identified during the period 15/05/2013-06/07/2014. 18 randomly sampled dropouts were manually verified and were found to be correctly identified. It was therefore assumed that the algorithm could detect sensor dropouts accurately.

Through the algorithmic identification, it was found that data dropouts occurred with a relatively high frequency over a wide range of timescales. Dropout periods lasting less than 5 minutes can be neglected for most analysis as their effect is to aggregate measured energy use for just a short period into the future. However, longer dropouts that offset large, aggregated readings to the next day, week or month can significantly affect the quality of analysis. The algorithm was adjusted to distinguish between 5-minute gaps and longer gaps. The 5-minute gaps were corrected by splitting the peak value over the previous two time slots. Gaps larger than five minutes and their peaks were deleted from the data for the short-time-dependent analysis in Section 4.2.5, so that unusually high peaks did not affect average energy use during occupancy.

4.2.3 Inferring Occupancy by CO_2-PIR Correlation

CO_2 data was available for two of the houses for a limited time period. As it has been shown in previous studies that the motion sensor data alone cannot provide accurate readings of the number of people [65], it was investigated whether CO_2 could be used to provide any further context. This investigation began with a visual inspection into the trends seen with the motion sensor data across houses C and G.

4.2.3.1 House C

The visual correlation between CO_2 and PIR readings in house C largely followed a logical pattern: the greatest daily peak in PIR readings corresponded to the greatest daily peak in CO_2, with a delay of around 0-2 hours. This pattern is illustrated with a sample of the available data in Figure 4-4. The delay between peak values was significantly longer than the 10-20 minutes observed in previous studies using room-level CO_2 sensing [65], likely due to the time required for a localised increase in CO_2 level to circulate to the single CO_2 sensor in the house.

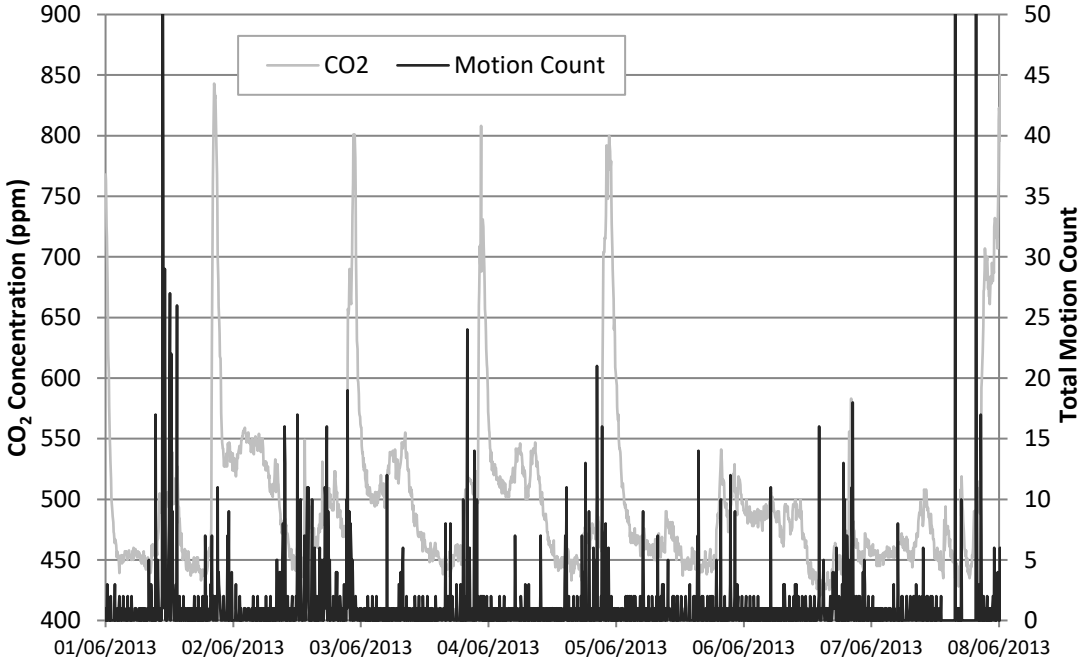


Figure 4-4 - Comparison of daily peak of total motion sensor and CO_2 readings for House C over one week period

4.2.3.2 House G

The visual correlation between CO_2 and motion sensor data for house G, illustrated in Figure 4-5, was less well defined. This may be due to the higher variation in daily cycle of occupancy or the lack of clearly defined daily peak. This may also indicate

the importance of the location of a single-point CO_2 sensor: if a large number of occupants are present in a room far away from the sensor, depending on the layout of the house, the local peak in CO_2 level may not be recorded before it is dispersed or ventilated to the outside.

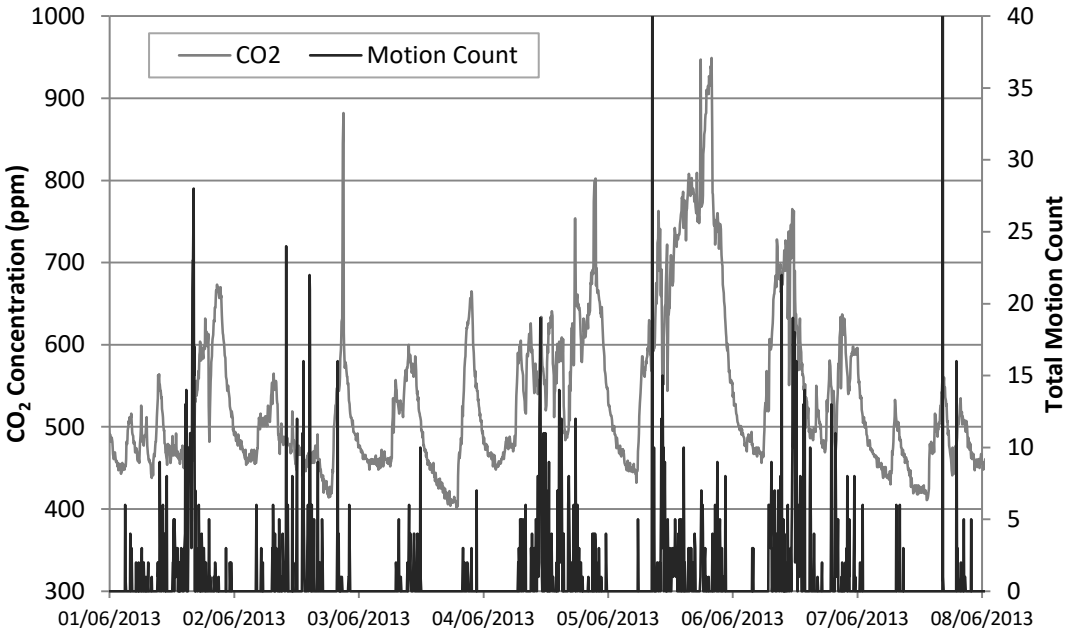


Figure 4-5 - Comparison of daily peak of total motion sensor and CO_2 readings for House G over one week period

4.2.4 Correlation of PIR/ CO_2 with energy use

Investigation into how the energy use relied on occupancy in the Green Street buildings began with a comparison of the direct correlation between various energy measurements and the magnitude of motion sensor or CO_2 readings. This provided some insight into how well the PIR and CO_2 data related to actual occupancy levels as well as with energy use.

It was expected that for most of the energy measurements, there would be a high number of zero-energy readings for all levels of occupancy. This is because an occupant does not use all forms of energy in a house all the time s/he is present. The expected indicator of good correlation was that the highest cases of energy use occurred during high PIR or CO_2 readings.

Figure 4-6 shows a comparison of Daily (Figures a and d), Hourly (b and e) and 5-minute (c and f) occupancy measurements relative to imported electricity for the period 15/05/2013 – 06/07/2014. It can be seen that on a daily level, a positive

correlation between detected occupancy and electricity demand exists: higher demands generally occurred on days with higher total motion counts and higher average CO_2 levels. However, on an hourly and 5-minute scale this correlation became less distinct. No correlation at shorter time scales could be observed with other energy uses, which has several possible reasons and implications:

- Given that motion and CO_2 sensors are centrally located, if the occupant is out of sensor range inside a room and using energy, their presence will not be logged until they enter/leave – this creates a disconnect between short-term occupancy measurement and energy use.
- Motion sensor count and CO_2 level cannot be directly used to infer the number of active occupants at a given time, but when summed/averaged over a day inaccuracies are evened out allowing for better correlation with daily energy. It should be noted that the number of active occupants does not directly imply a proportional increase in energy use, as behavioural aspects and external factors have a significant influence on individual energy use, however the observed positive correlation on a daily scale suggests that, on average, there is some relationship present between number of occupants and energy use.
- The disconnect may also imply energy waste, if true occupant presence is not related to short-term energy use, this implies that energy consumed is not always providing benefit to the occupants and could be saved.

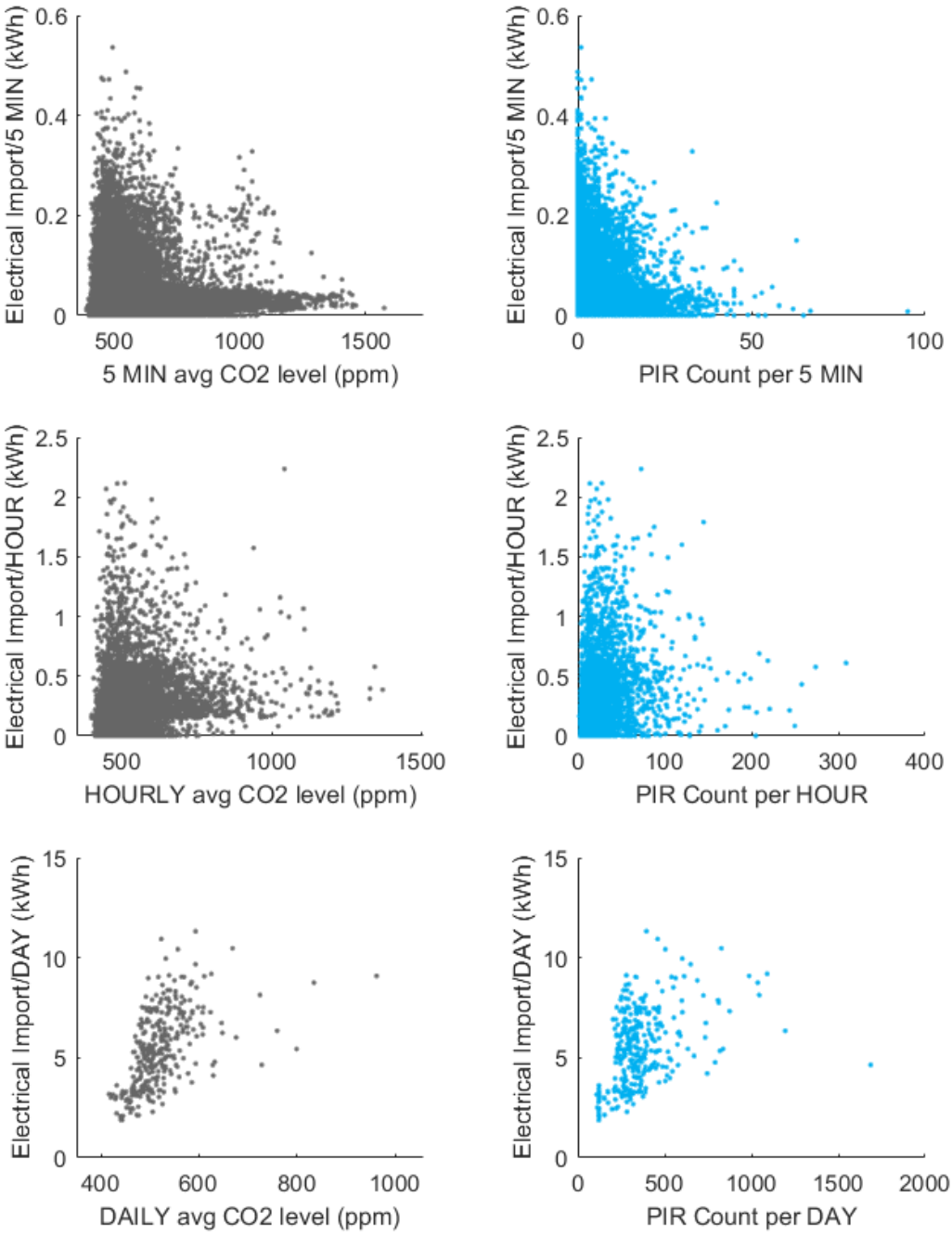


Figure 4-6 - House C. Correlation between Energy and Daily, Hourly and 5 Minute values for CO2 and PIR, using data 15/05/13-06/07/14

As daily values showed the strongest observable relationship to energy use, these were used to compare the relative relationship of various domestic energy uses to the motion count and CO_2 concentration data in houses C and G.

Most energy uses in house C (Figure 4-7) show the pattern that the lowest daily consumption occurs on days with the lowest motion count and CO_2 concentration – and so by implication the lowest number of occupants present. The strongest trends appear to have a linear pattern, allowing a correlation coefficient to be calculated, as

shown in Table 4-2. Here, a coefficient close to zero implies no correlation between two variables, and a coefficient closer to -1 or 1 implies a strong negative or positive correlation, respectively. For each correlation coefficient, the corresponding 'p-value' is also shown in Table 4-3. A low p-value implies there is a low probability of the correlation coefficient occurring when no relationship exists between the variables. It is therefore expected that a strong correlation between variables would have a high correlation coefficient and low p-value.

The strongest correlation of energy with occupancy in house C was seen for water, hot water, downstairs sockets and electrical import, all of which showed an approximately linear positive trend. This was confirmed by the correlation coefficient (Table 4-2) and the low probability of observing this trend if no relationship existed (P-value shown in Table 4-3). The trend suggests that the days with the highest motion count and highest average CO_2 levels are those with the greatest consumption, with two potential implications:

- Electrical and water consumption depend on the number of people present
- Electrical and water consumption depend on the length of time the house is occupied during the day

Without explicit knowledge of the number of people in the house at a given time, it is difficult to confirm which of these implications has the strongest influence.

The weakest correlation in house C is seen in the extractor and upstairs sockets, both of which appear to be independent of occupancy levels. In the case of upstairs sockets, it appears that the building occupants have a small, constant load, but do not often use much more electrical energy upstairs.

All other energy uses show a weak positive relationship with occupancy. The low p-values in Table 4-3 imply that the weak correlation is not coincidental, but that occupancy rate was not a significant factor in the magnitude of energy used, suggesting either:

- Dependence on other factors - for example, the energy used by the heating system will heavily depend on outside temperature and so is difficult to relate only to occupancy.

- Potential indication of energy use without full benefit to the occupant, resulting in wasted energy. In the case where higher levels of energy are being used while occupants are not present, this may indicate that systems or appliances have been left on when not needed. However, this is difficult to prove conclusively with the evidence available.

The correlation between the gas/heating use and the CO_2 data was significantly different to that for the motion data, as shown in Table 4-2 and Table 4-4. The reasons for this were further explored in Section 4.2.5.

The energy consumption of house G follows some similar trends to C, as shown in Figure 4-8, Table 4-4 and Table 4-5. Once again, the strongest correlation between energy and occupancy was found with electrical and water consumption. The heat recovery system was independent of occupancy and other energy uses show a weak correlation. It should be noted that in house G, the average CO_2 level correlated more strongly than the motion count for several energy measurements including electrical import, downstairs socket and cooker use. This may be due to the CO_2 sensor being placed closer to commonly used electrical equipment in house G, where the kitchen and living room are on the same storey, while in house C the kitchen is located one storey below the CO_2 sensor. It may also imply that the motion sensor location/sensitivity in house G was not as effective at detecting true occupancy as in house C.

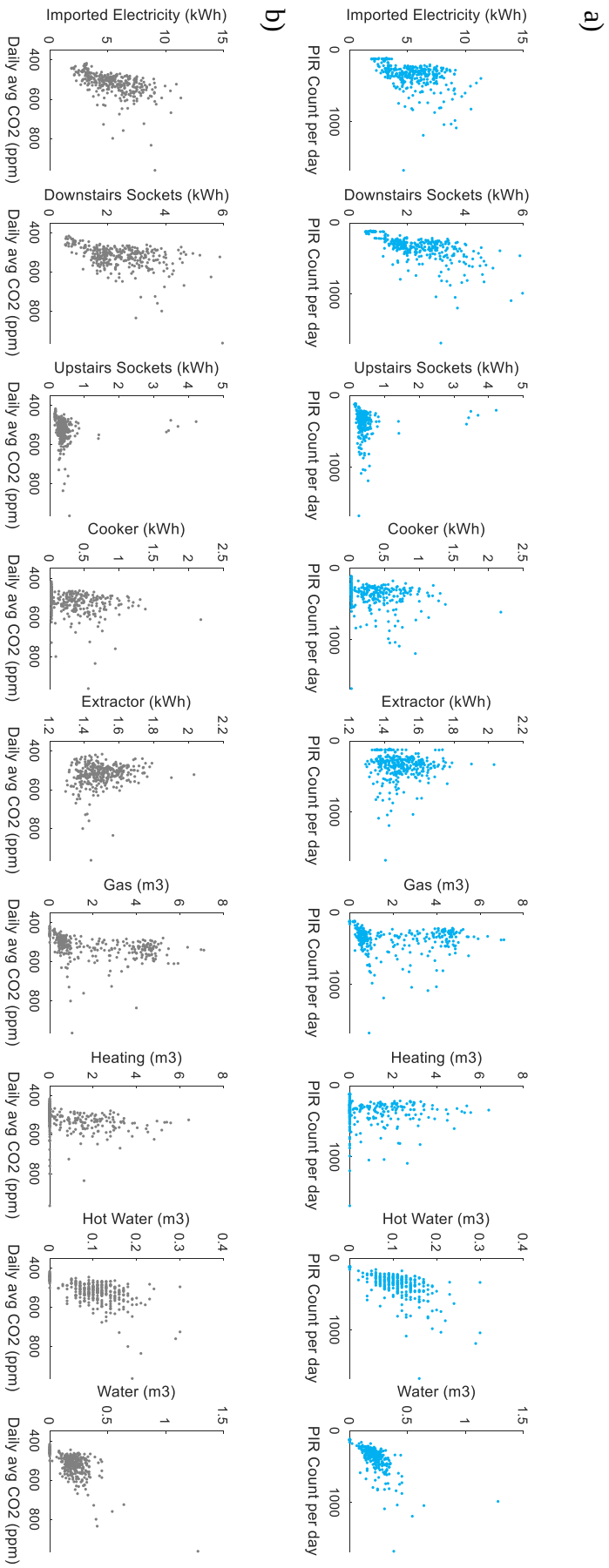


Figure 4-7 - House C. Correlation between energy use and a) motion count b) CO_2 measurements

Table 4-2 - House C. Correlation Coefficient between Daily Sensor Measurements (see Table 4-1 for sensor details)

	CO_2	Total PIR	PIR Down	PIR Up	Cooker	Extractor	Gas	Heating	Elec. Export	Elec. Import	Sockets Down	Sockets Up	Water	Hot Water		AVERAGE
CO_2	1.00	0.69	0.63	0.67	0.27	-0.16	0.35	0.29	0.02	0.53	0.45	0.06	0.66	0.54		0.38
Tot. PIR	0.69	1.00	0.92	0.96	0.29	-0.02	0.14	0.08	0.12	0.43	0.59	0.03	0.71	0.64		0.43
PIR Down	0.63	0.92	1.00	0.78	0.31	-0.01	0.18	0.11	0.14	0.45	0.59	0.04	0.70	0.67		0.42
PIR Up	0.67	0.96	0.78	1.00	0.25	-0.02	0.11	0.05	0.09	0.38	0.53	0.02	0.64	0.55		0.39
Cooker	0.27	0.29	0.31	0.25	1.00	0.12	0.17	0.08	0.05	0.49	0.41	0.03	0.39	0.48		0.26
Extractor	-0.16	-0.02	-0.01	-0.02	0.12	1.00	0.03	-0.07	0.09	0.02	0.08	-0.02	0.06	0.05		0.01
Gas	0.35	0.14	0.18	0.11	0.17	0.03	1.00	0.80	0.06	0.67	0.17	-0.01	0.17	0.29		0.24
Heating	0.29	0.08	0.11	0.05	0.08	-0.07	0.80	1.00	0.07	0.57	0.12	-0.03	0.05	0.12		0.17
Elec. Export	0.02	0.12	0.14	0.09	0.05	0.09	0.06	0.07	1.00	0.09	0.10	0.05	0.06	0.04		0.08
Elec. Import	0.53	0.43	0.45	0.38	0.49	0.02	0.67	0.57	0.09	1.00	0.65	0.19	0.48	0.51		0.42
Sockets D	0.45	0.59	0.59	0.53	0.41	0.08	0.17	0.12	0.10	0.65	1.00	0.10	0.68	0.57		0.39
Sockets U	0.06	0.03	0.04	0.02	0.03	-0.02	-0.01	-0.03	0.05	0.19	0.10	1.00	0.07	0.07		0.05
Water	0.66	0.71	0.70	0.64	0.39	0.06	0.17	0.05	0.06	0.48	0.68	0.07	1.00	0.83		0.42
Hot Water	0.54	0.64	0.67	0.55	0.48	0.05	0.29	0.12	0.04	0.51	0.57	0.07	0.83	1.00		0.41

Table 4-3 - House C. P-values for correlation coefficients

	CO_2	Total PIR	PIR Down	PIR Up	Cooker	Extractor	Gas	Heating	Elec. Export	Elec. Import	Sockets Down	Sockets Up	Water	Hot Water
CO_2	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.71	0.00	0.00	0.29	0.00	0.00
Tot. PIR	0.00	1.00	0.00	0.00	0.00	0.71	0.01	0.15	0.03	0.00	0.00	0.61	0.00	0.00
PIR Down	0.00	0.00	1.00	0.00	0.00	0.80	0.00	0.05	0.01	0.00	0.00	0.52	0.00	0.00
PIR Up	0.00	0.00	0.00	1.00	0.00	0.68	0.06	0.35	0.09	0.00	0.00	0.71	0.00	0.00
Cooker	0.00	0.00	0.00	0.00	1.00	0.03	0.00	0.15	0.33	0.00	0.00	0.63	0.00	0.00
Extractor	0.00	0.71	0.80	0.68	0.03	1.00	0.61	0.22	0.10	0.77	0.13	0.75	0.27	0.41
Gas	0.00	0.01	0.00	0.06	0.00	0.61	1.00	0.00	0.30	0.00	0.00	0.83	0.00	0.00
Heating	0.00	0.15	0.05	0.35	0.15	0.22	0.00	1.00	0.23	0.00	0.04	0.61	0.36	0.03
Elec. Export	0.71	0.03	0.01	0.09	0.33	0.10	0.30	0.23	1.00	0.11	0.06	0.33	0.29	0.45
Elec. Import	0.00	0.00	0.00	0.00	0.00	0.77	0.00	0.00	0.11	1.00	0.00	0.00	0.00	0.00
Sockets D	0.00	0.00	0.00	0.00	0.00	0.13	0.00	0.04	0.06	0.00	1.00	0.08	0.00	0.00
Sockets U	0.29	0.61	0.52	0.71	0.63	0.75	0.83	0.61	0.33	0.00	0.08	1.00	0.19	0.21
Water	0.00	0.00	0.00	0.00	0.00	0.27	0.00	0.36	0.29	0.00	0.00	0.19	1.00	0.00
Hot Water	0.00	0.00	0.00	0.00	0.00	0.41	0.00	0.03	0.45	0.00	0.00	0.21	0.00	1.00

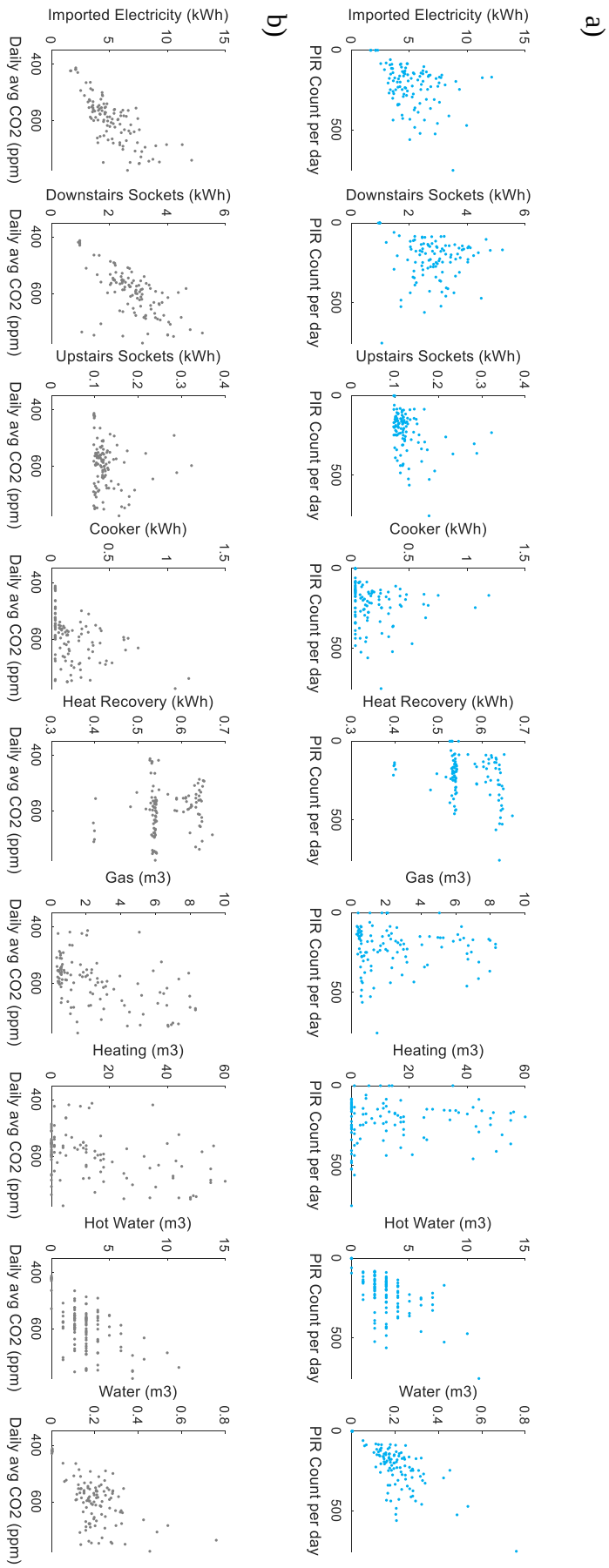


Figure 4-8 - House G. Correlation between energy use and a) motion count b) CO_2 measurements

Table 4-4 - House G. Correlation Coefficients between Sensor Measurements (see Table 4-1 for sensor details)

	CO_2	Total PIR	PIR Down	PIR Up	Cooker	Extractor	Gas	Heating	Elec. Export	Elec. Import	Sockets Down	Sockets Up	Water	Hot Water		AVERAGE
CO_2	1.00	0.43	0.49	0.28	0.41	-0.16	0.47	0.43	-0.24	0.76	0.64	0.12	0.55	0.54		0.36
Tot. PIR	0.43	1.00	0.94	0.90	0.12	0.31	0.01	-0.05	0.17	0.38	0.10	0.34	0.70	0.59		0.38
PIR Down	0.49	0.94	1.00	0.71	0.09	0.23	0.06	0.01	0.15	0.39	0.16	0.25	0.65	0.54		0.36
PIR Up	0.28	0.90	0.71	1.00	0.12	0.36	-0.07	-0.12	0.18	0.29	0.00	0.40	0.64	0.54		0.33
Cooker	0.41	0.12	0.09	0.12	1.00	-0.10	0.16	0.13	-0.12	0.53	0.44	0.16	0.29	0.27		0.19
Extractor	-0.16	0.31	0.23	0.36	-0.10	1.00	-0.56	-0.57	0.43	-0.22	-0.18	0.25	0.25	0.13		0.01
Gas	0.47	0.01	0.06	-0.07	0.16	-0.56	1.00	1.00	-0.55	0.60	0.36	-0.12	0.02	0.21		0.12
Heating	0.43	-0.05	0.01	-0.12	0.13	-0.57	1.00	1.00	-0.56	0.56	0.34	-0.15	-0.05	0.13		0.09
Elec. Export	-0.24	0.17	0.15	0.18	-0.12	0.43	-0.55	-0.56	1.00	-0.45	-0.27	0.28	0.24	0.02		-0.06
Elec. Import	0.76	0.38	0.39	0.29	0.53	-0.22	0.60	0.56	-0.45	1.00	0.68	0.03	0.49	0.50		0.35
Sockets D	0.64	0.10	0.16	0.00	0.44	-0.18	0.36	0.34	-0.27	0.68	1.00	-0.06	0.18	0.14		0.19
Sockets U	0.12	0.34	0.25	0.40	0.16	0.25	-0.12	-0.15	0.28	0.03	-0.06	1.00	0.32	0.29		0.16
Water	0.55	0.70	0.65	0.64	0.29	0.25	0.02	-0.05	0.24	0.49	0.18	0.32	1.00	0.84		0.39
Hot Water	0.54	0.59	0.54	0.54	0.27	0.13	0.21	0.13	0.02	0.50	0.14	0.29	0.84	1.00		0.37

Table 4-5 - House G. P-Values for correlation coefficients

	CO_2	Total PIR	PIR Down	PIR Up	Cooker	Extractor	Gas	Heating	Elec. Export	Elec. Import	Sockets Down	Sockets Up	Water	Hot Water
CO_2	1.00	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.01	0.00	0.00	0.21	0.00	0.00
Tot. PIR	0.00	1.00	0.00	0.00	0.23	0.00	0.94	0.63	0.07	0.00	0.30	0.00	0.00	0.00
PIR Down	0.00	0.00	1.00	0.00	0.33	0.02	0.52	0.89	0.12	0.00	0.10	0.01	0.00	0.00
PIR Up	0.00	0.00	0.00	1.00	0.20	0.00	0.50	0.23	0.07	0.00	0.96	0.00	0.00	0.00
Cooker	0.00	0.23	0.33	0.20	1.00	0.31	0.10	0.16	0.20	0.00	0.00	0.10	0.00	0.00
Extractor	0.10	0.00	0.02	0.00	0.31	1.00	0.00	0.00	0.00	0.02	0.06	0.01	0.01	0.18
Gas	0.00	0.94	0.52	0.50	0.10	0.00	1.00	0.00	0.00	0.00	0.00	0.22	0.81	0.02
Heating	0.00	0.63	0.89	0.23	0.16	0.00	0.00	1.00	0.00	0.00	0.00	0.13	0.60	0.16
Elec. Export	0.01	0.07	0.12	0.07	0.20	0.00	0.00	0.00	1.00	0.00	0.01	0.00	0.01	0.82
Elec. Import	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	1.00	0.00	0.78	0.00	0.00
Sockets D	0.00	0.30	0.10	0.96	0.00	0.06	0.00	0.00	0.01	0.00	1.00	0.52	0.07	0.15
Sockets U	0.21	0.00	0.01	0.00	0.10	0.01	0.22	0.13	0.00	0.78	0.52	1.00	0.00	0.00
Water	0.00	0.00	0.00	0.00	0.00	0.01	0.81	0.60	0.01	0.00	0.07	0.00	1.00	0.00
Hot Water	0.00	0.00	0.00	0.00	0.00	0.18	0.02	0.16	0.82	0.00	0.15	0.00	0.00	1.00

4.2.5 Using PIR and CO_2 to indicate binary occupancy

Without some ground truth data to provide context, it is impossible to accurately tell how many people are present from PIR and CO_2 data alone. As investigated above, it may be possible to use localised CO_2 levels to strengthen the PIR's reliability and infer occupancy levels, although this is extremely difficult to verify without known accurate occupancy measurement for at least some period of time in order to train a model.

As an alternative measure, this section covers the exploration of using PIR and CO_2 data to infer binary occupancy – whether occupants are present or not at any given time. Once again, it is not possible to verify without true occupancy data, but an estimate of effectiveness can be made by comparing different methods and correlation to energy use. A test was conducted on the five-minute resolution data by comparing the average energy use when assumed unoccupied to when assumed occupied: here a greater difference implies the means of assuming occupant presence was more successful. Various assumptions were applied to allow binary occupancy to be estimated using the PIR and CO_2 level data:

Assumptions for PIR:

- A nonzero value of motion sensor count for either upstairs or downstairs denotes occupant presence for that five-minute timestep.
- Zero values for both upstairs and downstairs denote occupant absence for that timestep.
- Downstairs sockets are reliant on downstairs PIR only, and upstairs sockets on upstairs PIR.
- The kitchen cooker/hob circuit is reliant on downstairs PIR only, assuming that the occupant must be in or near the kitchen to use the hob.

Assumptions for three methods of classifying binary occupancy through CO_2:

1. Occupancy is assumed when CO_2 is above a certain threshold in the given five-minute timestep – in this case above the mean value over the observed period (519ppm for house C and 597ppm for G).
2. Occupancy is assumed when current CO_2 is higher than the mean over the previous 30 minutes.

3. Occupancy is assumed when the mean CO_2 over the last 10 minutes is higher than the mean over the previous 1 hour – this was tested to reduce the effect of minor variations caused by noise on the sensor data.

Figure 4-9 shows the results of this analysis for house C using PIR motion sensors. Graphs with both occupied and non-occupied values close to the overall average (denoted by a grey dashed line) show a lower correlation between assumed occupancy and energy use, while a larger percentage difference between occupied and non-occupied suggests a higher correlation. It should be noted that the mean occupied value does not represent an estimated value of energy in use, as people will not use all systems all the time they are present.

The percentage difference for all occupancy estimation methods is summarised in Figure 4-10 for house C and Figure 4-11 for house G. It can be seen that no single method of assuming occupancy consistently correlated best with all forms of energy use. PIR was generally better correlated to water and cooker use, while CO_2 was better correlated to heating. This implies that either:

- The relative location of the motion and CO_2 sensors allowed each to detect some activities better than the other. Referring to the house layouts (Appendix 10.1), some logical connections can be seen. For example, both houses had PIR sensors directly outside the main bathroom and kitchen, meaning that when an occupant travels to these rooms to use water, the PIR was more likely to sense their presence than the living-room-based CO_2 sensor.
- Outside factors can bias the effectiveness of PIR or CO_2 in certain situations – for example, CO_2 is correlated to heating use much more strongly than PIR. This may be because occupants are more likely to keep windows closed during winter when heating is most required, so conserving high CO_2 levels.

One of the possible explanations for the correlation between CO_2 level and heating use is that the users were more likely to keep their windows closed during the winter period, allowing CO_2 levels to increase and thus associating higher CO_2 with heating. However, the presence of the constantly-running ventilation system in the houses could negate this effect and should theoretically keep the CO_2 level constant. In order to investigate this further, the average CO_2 level was calculated for the summer (1st

April – 30th September) and winter (1st October – 30th March) periods within the available data.

As presented in Table 4-6, for both of the houses fitted with CO_2 sensors the average CO_2 level is higher during the heating period. The difference is much larger in House G, although it should be noted that the data set for this house was smaller, and in particular included only 18 days of data for the heating season, whereas 182 days were available for House C. However, the trend towards higher CO_2 levels in general during the heating period supports the hypothesis that CO_2 was conserved by keeping windows closed during colder weather.

Table 4-6 - Average CO_2 levels for the summer and winter periods

	Cooling period (summer)	**Heating period (winter)**
House C	512 ppm	530 ppm
House G	580 ppm	696 ppm

The most consistent strong correlation for both the PIR and CO2 presence assumption methods was found with hot water and cooker use – this is logical as most domestic systems do not allow for these to be operated when the occupant is not directly controlling them.

The least correlation with estimated presence for both houses was found with the extraction ventilation system and upstairs sockets. In particular, the extraction system showed no correlation at all with occupancy. This potentially means that energy was being wasted while the houses were not occupied. It should be noted that slow-response systems such as heating/ventilation may need to run when occupants are not there to achieve comfort, but it is unlikely that they are required to run constantly.

The graphs also confirm that neither PIR nor CO_2 can be used to perfectly estimate whether occupants are present or away. The 'unoccupied' mean for water and cooker use was not zero, despite the assumption that occupants will not use either if not present in the building. These are 'false negative' occupancy estimations - reading zero occupancy when in reality there are occupants present. As there were only two PIR sensors and one CO_2 sensor per house, this can be explained by people being out of range of the sensors but still present in the house. This is further supported by both water use and cooking occurring in specific rooms, meaning that the occupants

directly involved in cooking or water use would not have been triggering the centrally-placed PIR/ CO_2.

With the data available it was impossible to prove if 'false positives' (reading occupancy when nobody is present) occurred, as it might be that occupants were present in the house without actively using any energy.

The high uncertainty and proven presence of sensing errors show that a system based purely on untrained analysis of limited CO_2 and PIR data cannot provide enough reliable occupant data to control energy services.

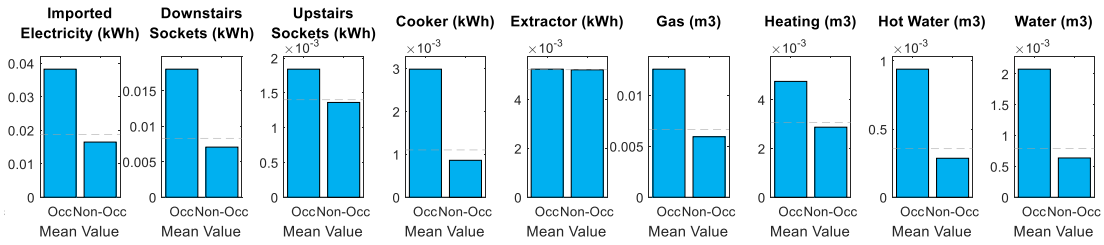


Figure 4-9 - Average 5-minute values for House C energy uses when occupied/not occupied, using motion sensor activity

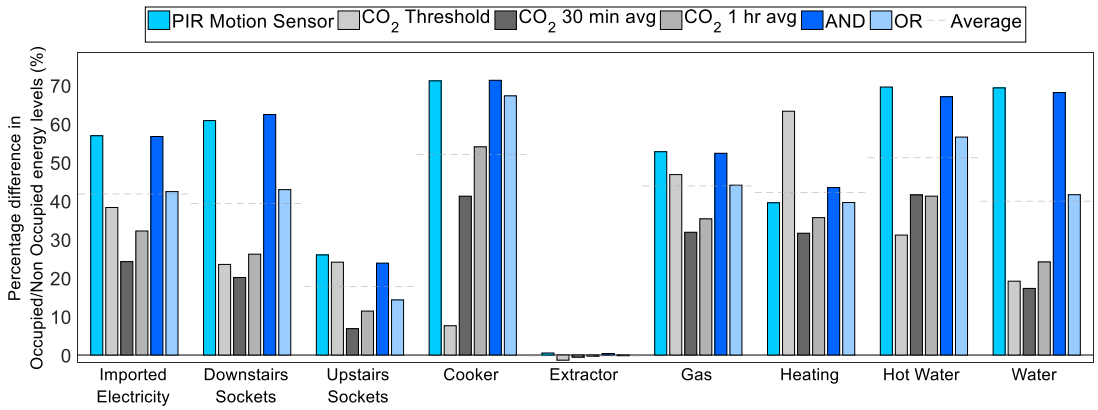


Figure 4-10 - Percentage Increase in Energy Use when Occupied, Comparison of various Occupancy Estimation Methods for House C

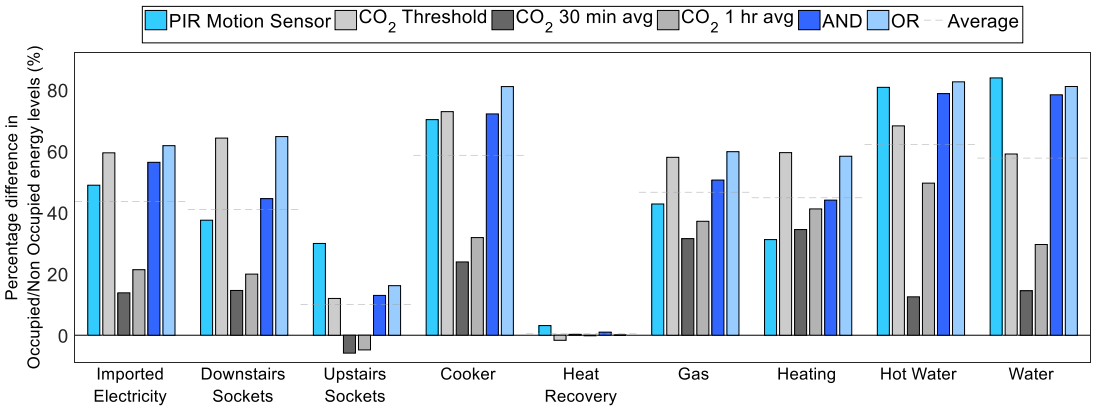


Figure 4-11 - Percentage Increase in Energy Use when Occupied, Comparison of various Occupancy Estimation Methods for House G

4.2.6 Low Occupancy Periods - Extended occupant absence during the heating period

As the previous sections show, it was difficult to verify when exactly the occupants were not present in the houses from a limited number of motion and CO_2 sensors. One instance in which it can be confidently said that the house is not occupied is during extended periods of absence: when the occupants are on holiday, for example.

It was carefully considered which combinations of parameters denote low or zero occupancy. This study was intended to investigate the relationship between energy use and occupancy; it could not be assumed therefore that energy uses such as socket loads and heating use would be consistently low when occupants were not present, as it is feasible that they would be left switched on during periods of absence. For the purpose of this study, it was assumed that:

- Typical domestic water outlets are not possible to use when occupants are not present in the home.
- If all sensors and meters read zero, the cause is likely dropout and not occupant absence, as it has been confirmed in Section 4.2.2 that prolonged dropouts occurred in this dataset.

Water consumption was therefore used to verify the motion sensor readings denoting extended periods of occupant absence. Two examples of absence during the heating period (October to April) were identified in houses C and D: 22/12/2013-03/01/2014 for house C and 04/01/2014-12/01/2014 for house D. Thus the impact of different heating behaviours was investigated.

Figure 4-12 and Figure 4-13 show water, electrical, heating and ventilation use over the two identified periods of absence. For each house, the consistent drop in both water/electric use confirmed absence over the period, as detailed above. In house C, the motion sensor registered a consistent 1-2 readings throughout the period, suggesting that the sensor was too sensitive or was triggered by movement outside the building through windows. The extractor was in near-constant operation over the whole absence period – it was not affected by occupancy level, presumaby supplying the same amount of fresh air to the house despite the fact that no occupants were present. This strengthens the conclusions made in Section 4.2.5 that the extract ventilation system was uncorrelated with occupant presence. A point of interest in Figure 4-12c) is that the heating was apparently turned off during the occupants' absence. This is not an automated behaviour and so must have been manually chosen by the occupants. A slight peak in heating demand can be seen after the off-period, corresponding to the need to heat the house from a lower than typical temperature upon returning.

In house D, a shorter absence period can be seen & verified by water/electricity use. The motion sensor count is zero during absence, unlike house C, suggesting that this motion sensor was not affected by outside movement or air currents. The extractor was once again apparently unaffected by whether occupants were inside the house or not.

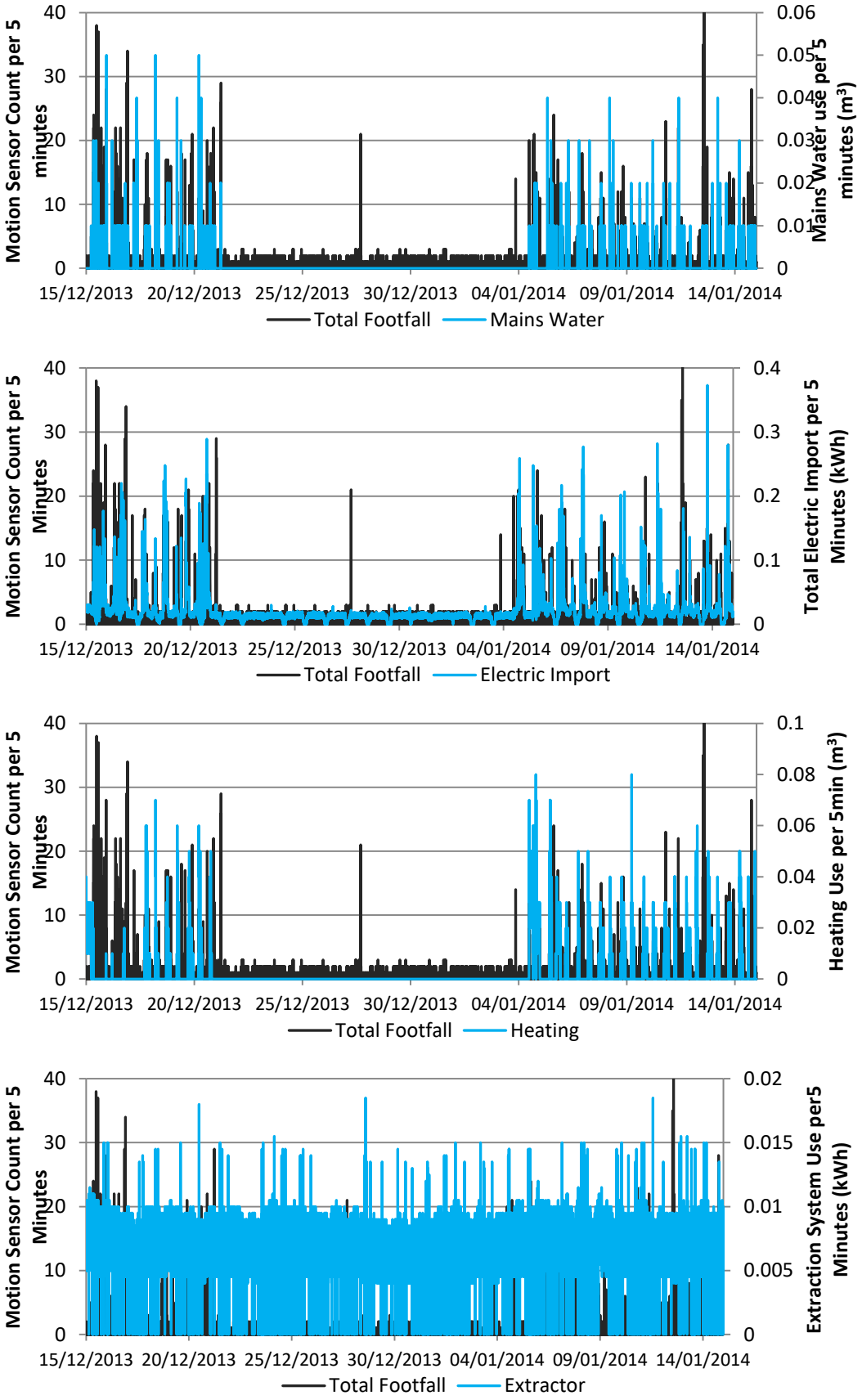


Figure 4-12 - a) Mains Water b) Electric Import c) Heating d) Extractor energy for House C during a period of occupant absence in the heating period

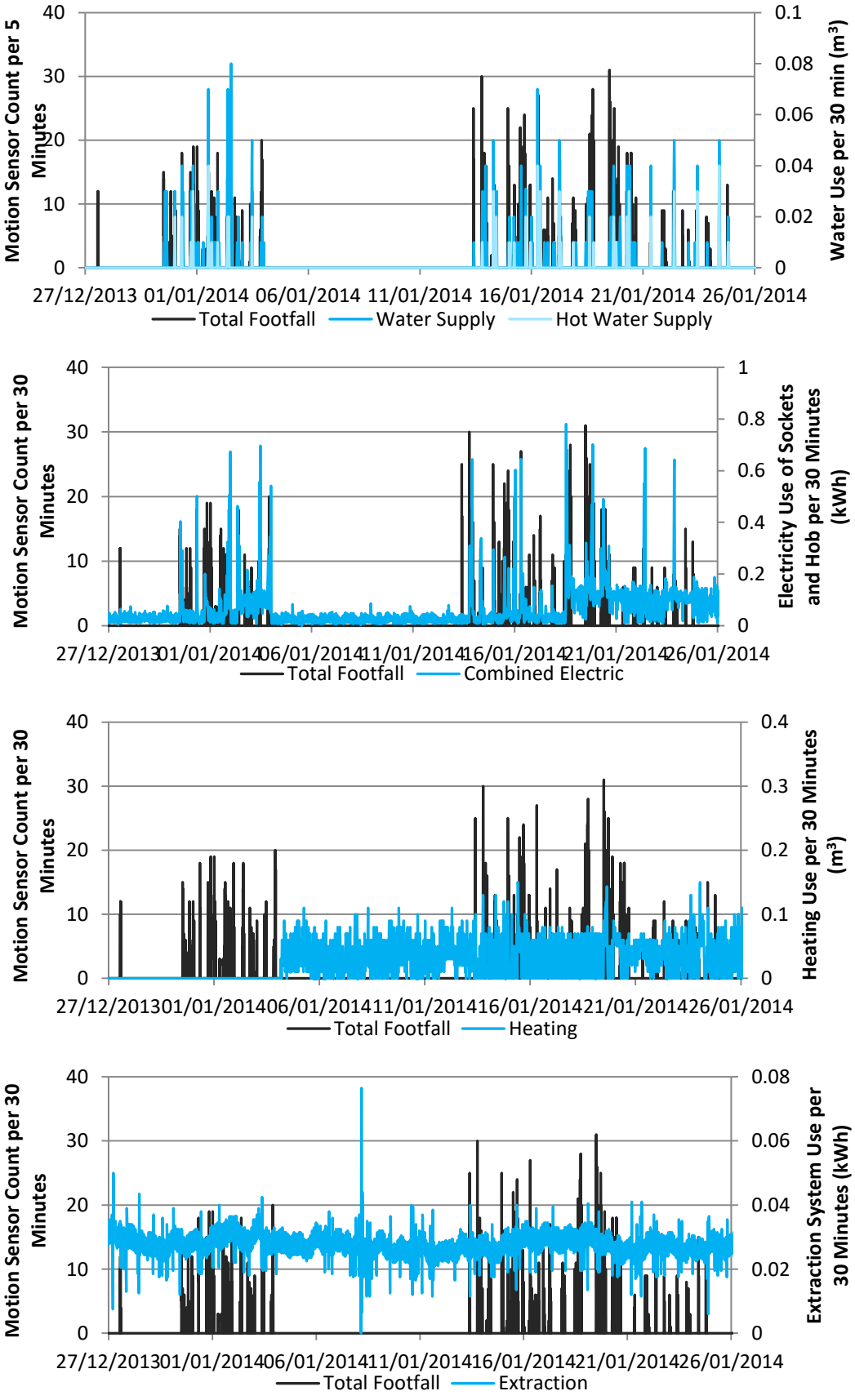


Figure 4-13 - a) Mains Water b) Electric Import c) Heating d) Extractor energy for House D during a period of occupant absence in the heating period

Heating was also consistent over the absence and the following week: the same amount of energy was used for heating when the occupants were not in the house. In total, the use of 12.92m^3 of natural gas was recorded during the period 04 Jan - 12 Jan. Assuming the energy derived from natural gas is 10.75kWh/m^3 [182] and a boiler efficiency of 90% [183], the heating energy delivered over the 9 day period was 125kWh – nearly 14kWh per day. It should be noted that retaining some level of home heating during an absence can be beneficial, in order to keep pipes from freezing, ensure the house is comfortable as soon as the occupants return and eliminate a spike in heating demand when occupants return home. However, the above heating behaviours show that without intervention, heating beyond this base level will continue to be supplied during extended periods of occupant absence, leading to a waste of energy in the home. Systems to automate the heating and ventilation within these homes based on real-time occupancy measurement would reduce the overall energy demand of the houses.

4.2.7 Discussion

One of the major conclusions to draw from the study of Green Street data is that gaining meaningful, reliable information can be extremely difficult given limited data. The PIR and CO_2 data showed some similar trends, but did not agree closely enough to confidently estimate occupancy at a given time. It is also impossible to relate the magnitude of sensor reading to a physical number of people without prior training with known true occupancy rates.

Whole-house occupancy is difficult to relate to energy, as presence in the house does not imply that any particular activity must be taking place. It is also impossible to prove that energy being consumed is actually providing utility to the user without context on where the user is at a given time. This makes it difficult to highlight energy waste at any time other than extended periods of absence.

It was found that some measured variables could be used to verify periods of occupant absence over the longer term – for example a lack of water use over a whole day implies the occupant is not at home. However, this cannot be extended to shorter timescales as it is possible for the occupant to be present for short periods without using water.

However, it was proven that an occupant's behaviour towards heating control can have a significant impact on the energy used while away for a period of several days. In a study between two Green Street houses, it was shown that approximately 14kWh more energy per day was used on heating an empty home depending on whether the occupants turned off their heating prior to a holiday during winter 2013/2014.

The study also highlighted that data quality issues can affect what conclusions can be drawn from data analysis. For example, the frequent sensor dropouts can look similar to periods of occupant absence, making it difficult to reliably draw conclusions on changing energy use with occupancy. Issues such as this will be common to many sensing applications and inference should be designed to be as robust as possible to sensor dropouts, interference etc.

4.3 Explore Innovation Park Office Dataset

4.3.1 Building layouts, uses and data types available

The Explore Innovation Park site in Worksop, UK, contains several buildings used for office space, factory space for concrete production etc. Some energy and occupancy data was made available from the office building on this site, providing a case study of a larger-scale office space. Energy data from the building's metering system was made available over a one year period from Nov 2014 to September 2015, with readings from the building's energy meters available at a 30 minute resolution. As data was recorded at the meter level, the electricity use recorded is not disaggregated to a highly localised level, but is described broadly by area as shown in Table 4-7. Occupancy data was collected over five of the recorded months from the office's access control system, which logged the time and card ID when any occupant holding an access card passed through a controlled entrance within the office. A diagram of the doors monitored by the access control system is included in Appendix 10.2.

It should be noted that the access control in this building did not monitor all internal doors, and most monitored doors provided one-way access control, meaning that occupants needed to scan their card to go into a space, but did not need to scan to leave the same space. This means that the data cannot be used to directly represent the number of people present, as it was not known how long an occupant stayed after they scanned in. However, other ways of querying the available data were tested to highlight trends in occupancy rates over time.

Table 4-7 - Description of Energy Metering Data Available at EIP Offices

Meter Name	Description	Units
GF Office	Covers electrical energy demand from all ground floor areas not on a separate meter	kWh
1F Office	Covers electrical energy demand from all first floor office areas	kWh
2F Office	Covers electrical energy demand from all second floor office areas	kWh
Kitchen	Electrical energy demand from ground floor kitchen area	kWh
Heat Pump	Energy used by heat pump system in office building	kWh
IT Room AC	Ground floor IT Room climate control system	kWh
Roof Switch-board	Covers demand from switchboard on roof of office building	kWh

4.3.2 Inferring occupancy rates from unique visitors per day

In order to gain a broad idea of how the office use changed over longer time periods, the number of unique occupants logged to either the main entrance or main exit reception doors per day was calculated from the access data. Figure 4-14 shows a summary of the 7-month period for which both occupancy and energy data was available. Both occupancy and energy use follow a weekly pattern with distinct decreases during weekends, although some gaps are present in the data. Occupancy was not obtained for the period 28th Nov-31st Dec 2014. Some issues were present in the energy meter readings from 14th-27th Mar 2015. These periods were therefore omitted from any further calculations.

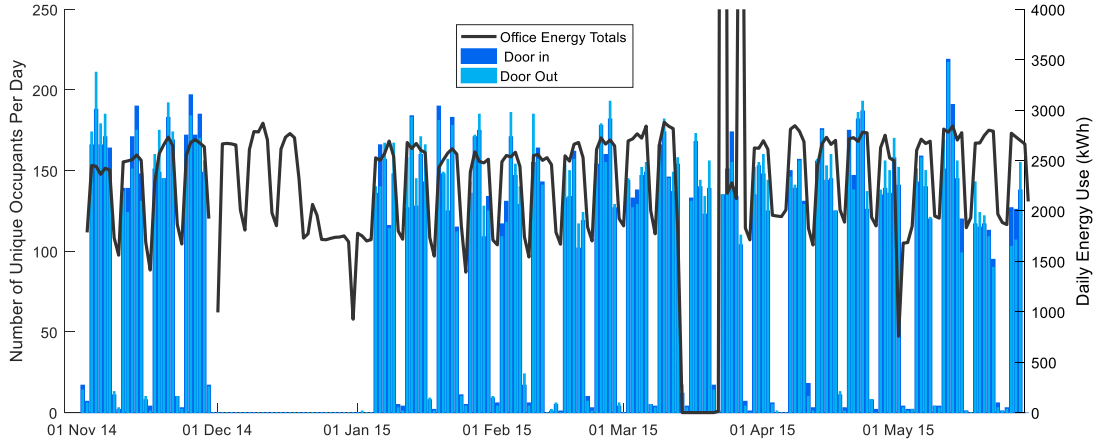


Figure 4-14 - Occupancy Rates vs Total Office Energy Use over a 7-Month Period

As the graph in Figure 4-14 shows, the reduction in total office energy use at weekends is smaller than the reduction in occupancy rates. This is to be expected to some extent, as some systems should be run at a background level regardless of the presence of occupants, giving a 'baseline' energy use for the building. However, the difference in magnitude may also indicate potential energy waste while the building is

close to unoccupied. A breakdown of the percentage decrease by energy meter is shown in Figure 4-15. It can be seen that the kitchen meter shows the greatest energy decrease at weekends: indicative of the lack of need to prepare food for working staff. The IT room air conditioning saw a slight increase in energy demand at weekends – it was expected that this area would have a consistent energy demand over the week, as the environment of any server equipment etc. should be kept constant. The slight increase may be due to a change in the temperature of the surrounding rooms: if occupied zones are conditioned to a lesser extent at weekends, the IT room may have to spend more energy to maintain its own temperature. The three floors of the office showed differing levels of variation from weekday to weekend. Unfortunately, as these meters covered both local heating systems as well as electrical demand from lighting etc, it is difficult to highlight where this difference comes from. It may be that the behaviours of occupants on some floors mean that more electrical equipment is left on over weekends, or that due to the layout of the building the relative heating demand varies floor to floor.

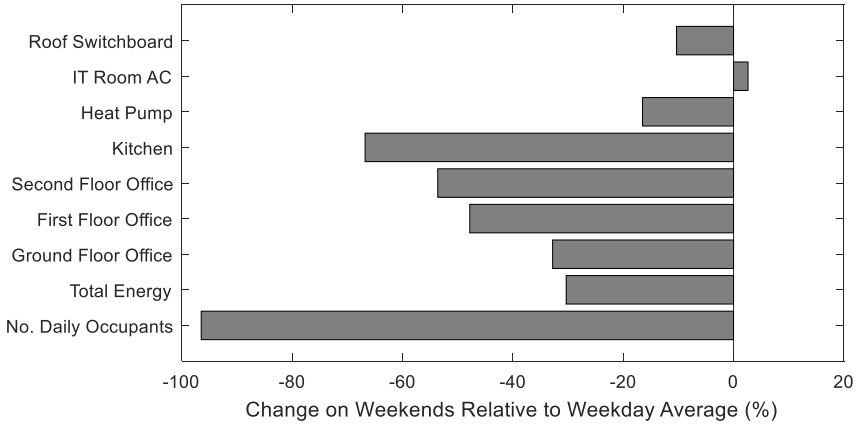


Figure 4-15 - Comparison of the Reduction in Occupancy Rates and Energy Use on Weekends relative to Weekdays

When constrained to weekdays only, the correlation coefficient between daily occupancy and energy rates shows some unexpected relationships. Table 4-8 and Table 4-9 indicate a relatively strong negative relationship between the heat pump energy demand and the number of occupants, suggesting that the presence of people producing heat in the office space offset some of the need for the heat pump during heating weather. The kitchen circuit showed a slight correlation to the number of occupants per day, which once again can be explained logically as higher numbers of occupants will require more services from the kitchen. However, the three floors of

the office showed significantly different relationships to the global number of occupants. The ground and second floors showed a positive relationship, indicating that a higher population on these floors generally meant a higher energy demand. On the first floor, however, little to no correlation was found. This has several potential implications:

- Significantly different use of the first floor in terms of what equipment uses the most energy – this option is unlikely, as the building floor plans show a similar layout for the first and second floors.
- Visitors to the first floor are not detected by the access system – given the similar use to the second floor, this again seems unlikely. It is possible that the higher number of meeting rooms on this floor mean that more visitors, who may bypass the access control system if accompanied by staff affect the energy use without being detected as occupants.
- Behavioural differences mean that equipment is left on regardless of occupancy on this floor.

Table 4-8 - Correlation Coefficient between Daily Occupancy and Energy Measurements

	Office Total	GF Office	1F Office	2F Office	Kitchen	Heat Pump	IT Room AC	Roof Switch-board
No. Daily Occupants	0.03	0.52	0.09	0.34	0.27	-0.58	-0.15	0.11

Table 4-9 – P-Value for Correlation between Daily Occupancy and Energy Measurements

	Office Total	GF Office	1F Office	2F Office	Kitchen	Heat Pump	IT Room AC	Roof Switch-board
No. Daily Occupants	0.68	0.00	0.22	0.00	0.00	0.00	0.04	0.12

4.3.3 Half-hourly Local Activity Levels

Although the exact number of people present in each location could not be known for the reasons discussed above, it was investigated whether a more localised measure of activity over the course of a day could be obtained using the number of events logged at internal doors. Each known door was assigned to a floor, as shown in the diagrams

in Appendix 10.2. A profile of the activity level of occupants could then be made, with a resolution of 30 minutes chosen to match the available energy data.

Figure 4-16 and Figure 4-17 show the mean daily energy profile for each of the energy meters plotted against the 30-minute activity level from the access control system, for weekdays and weekends respectively. The energy meters for the individual floors of the building were compared to the activity level of doors only from this floor, while the global energy meters were compared to the total activity over all monitored doors. The weekday activity level shows two distinct daily peaks at the start of the working day and at typical lunch break hours. The lack of a similarly sized peak at the end of the day is likely because most doors monitored only incoming occupants, although this may also indicate that leaving time was more variable between occupants. Weekends showed a much lower activity level, with a broader peak, likely as weekend occupants were not constrained to visiting for typical work hours.

Most meters saw an increase in energy use during occupied hours, with the exception of the IT room AC system, which was run at a constant rate independent of building occupancy. The individual floors appear to show a general trend of energy use following occupancy, with the consideration that sharp peaks in door activity denote more moving occupants, not necessarily more occupants overall. Interestingly, the second floor office shows a small dip in energy use during the spike of occupants leaving for lunch, while the first floor does not. This may indicate more wasteful energy behaviours of occupants on the first floor such as leaving equipment running, as discussed in the previous section. The heat pump use shows a daily peak that is offset by several hours from the daily peak in occupants. This may be due to high thermal mass in the building delaying the largest environmental conditioning loads. The kitchen also shows an offset daily peak, with the majority of energy used in the morning as occupants arrive and food is prepared for the busy lunchtime period.

Table 4-10 and Table 4-11 show the correlation coefficient and corresponding p-value between the half-hourly measured activity levels and the energy meter readings. The correlation of floor-level energy use and the associated floor's door activity is more strongly positive than the global number of occupants per zone. The heat pump use shows no correlation to activity, in contrast with the stronger negative correlation seen

in the daily values. The consistently low p-values show a low likelihood that the correlations could have been observed where no correlation existed, with the exception of the heat pump use and second floor activity level. As the correlation coefficient between these two was already small, this observation makes sense.

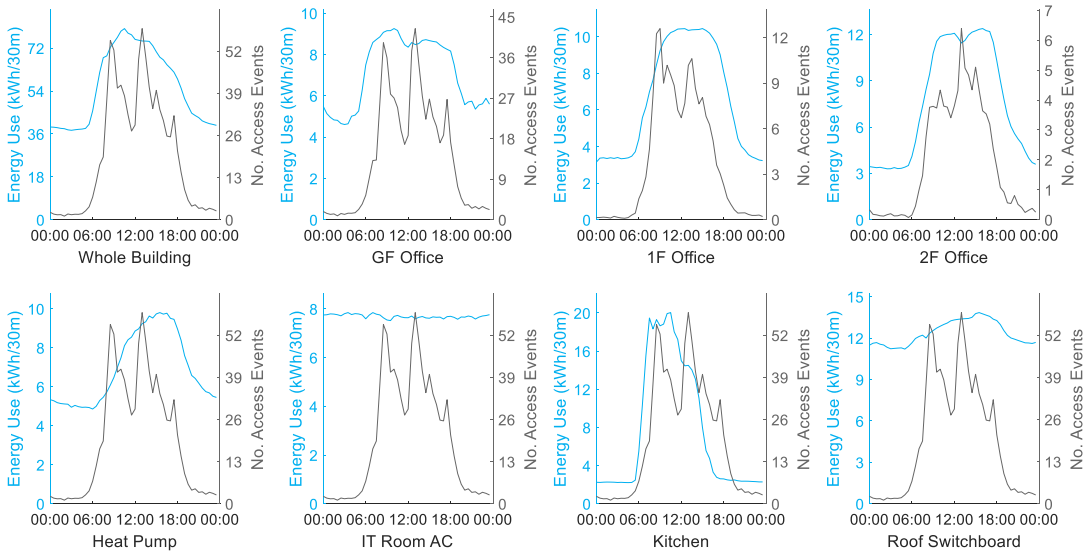


Figure 4-16 – Average Energy and Occupancy Profiles for each Office Meter – weekday

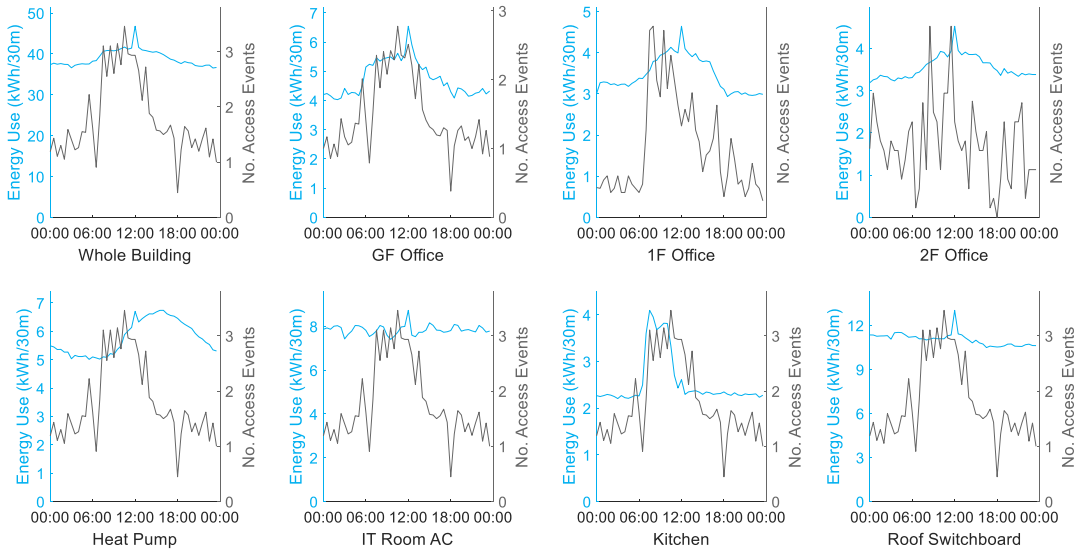


Figure 4-17 – Average Energy and Occupancy Profiles for each Office Meter – weekend

Table 4-10 - Correlation Coefficient between Half-Hourly Energy and Access Activity Level

	Roof Switch-board	Kitchen	IT Room AC	Heat Pump	GF Light Heat	1F Light Heat	2F Light Heat
Activity Total	0.16	0.49	-0.12	-0.10	0.52	0.52	0.55
Activity GF	0.18	0.51	-0.08	-0.06	0.48	0.53	0.55
Activity 1F	0.15	0.58	-0.09	-0.05	0.47	0.54	0.54
Activity 2F	0.19	0.46	-0.08	-0.02	0.45	0.54	0.55

Table 4-11 - P-Value for Correlation between Half-Hourly Energy and Access Activity Level

	Roof Switchboard	Kitchen	IT Room AC	Heat Pump	GF Light Heat	1F Light Heat	2F Light Heat
Activity Global	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Activity GF	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Activity 1F	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Activity 2F	0.0000	0.0000	0.0000	0.0525	0.0000	0.0000	0.0000

4.3.4 Discussion

From the analysis of daily totals, it appeared that the office energy use varied by less than occupancy rates on weekdays/weekends. Different floors showed different correlations to daily occupancy rates.

Energy profiles over the course of a day indicate some finer characteristics of the relationship between local occupancy and energy use, with some office areas appearing to be more conservatively managed than others. From this analysis, it became clearer that some systems were run continuously, and that some energy systems had an offset daily peak relative to the peak occupancy level.

As with the domestic case study, much of the potential to identify areas of energy waste relies on a more detailed picture of where occupants are in the building, in order to prove whether utility is actually being provided by expended energy. In the case of the office building, the energy meter values provided did not make a distinction between energy used for local lighting, heating and electrical demand, meaning that highlighting specific occupant behaviours that may be better improved or accounted for in the building controls was difficult. However, the characteristics identified in this study could be a useful starting point for further work making a deeper analysis

into the potential mismatches between energy use and occupant demand seen in some systems.

4.4 Dagenham Park School

Energy use data was also made available for Dagenham Park School, located in Dagenham, UK. In this case, no occupancy measurements were available, but the scheduled hours of occupancy for the school day were known. During its first year of operation, the building was known to exceed the contracted energy usage, and it was investigated internally whether this excess energy use was occurring during school hours or outside of school hours [184]. In this report, it was found that behavioural factors outside of scheduled building usage hours had a significant effect on the total electrical and gas energy use, as presented in Figure 4-18. It can be seen that the building performed under its electrical and biomass targets during 'core' school hours and 'plus' additional community opening hours, but that a significant percentage of the total energy use was expended during hours when the school was closed, for which time the energy use was not accounted for in the operational targets. It was concluded that poor energy behaviours such as leaving on small power and lighting loads overnight and during school holidays was the major contributing factor to the overall failure to achieve electrical energy targets during this year. The balance of biomass versus gas heating use was also attributed to behavioural factors, where it appeared the lower-carbon biomass system was not used by the school as much as was expected.

These results clearly highlight the likelihood of significant disparities between the way that a building and its systems are expected to be used and the results in practice. However, this study also shows a potential area for further study: without specifically querying the actual occupant use patterns of the building, it is not known whether the assigned core, additional and out-of-hours periods were appropriate. The higher than expected energy use during out-of-hours periods could indicate that the building was still in use during some of these hours, which would add further complexity to the issue of whether all energy use during this period was energy wasted. This is also true for the occupied hours: although performing under target, without specific knowledge of local occupancy it is impossible to identify whether all energy used in this period actually provided utility to building occupants, or whether the energy use during this time could be further reduced.

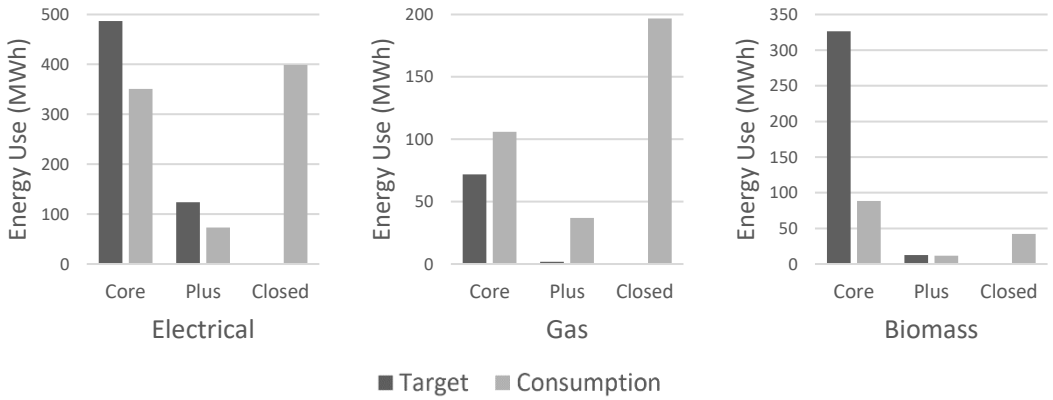


Figure 4-18 - Dagenham Park School Energy Use Breakdown over 1 Year Period

A qualitative assessment of energy behaviours in the school was made via site visit and interviews with staff. A tour of the building identified several areas where energy use was not being controlled according to the actual energy needs of occupants, with some examples presented in Figure 4-19. In these images it can be seen that electrical equipment within classrooms was left on despite occupants not being present, and that the electrical lighting in naturally lit circulation spaces within the school was in operation despite providing no benefit during daylight hours.

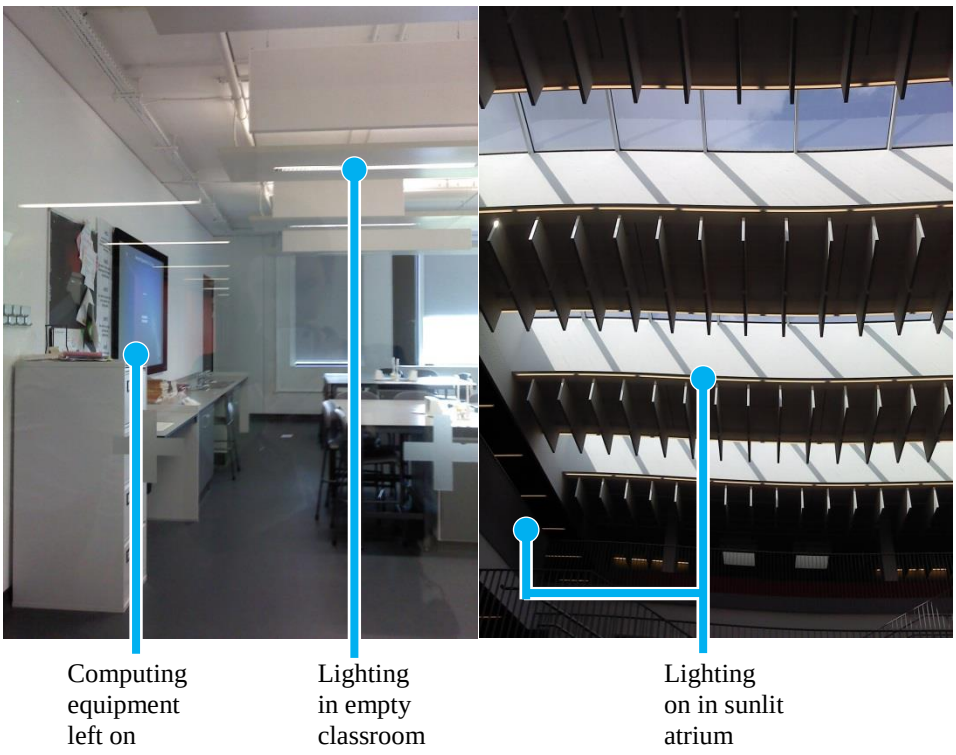


Figure 4-19 - Energy behaviours observed in Dagenham Park School

It was also found through the interview process that some areas of the school were left in excessive temperatures while occupied by classes of students, in particular

classrooms located on sun-facing sides of the building where a full class of students was present, which were served by a ventilation/AC system that could not address such localised peaks in temperature. This failure of the designed ventilation system highlights the need to consider the cooling load caused by large incoming groups of occupants. In the case of a school where room use by large groups of occupants should be centrally scheduled, the potential benefits of a pre-emptive control system based on the anticipated occupancy levels can be seen. The issue also highlights an important factor for occupant-centric control of services: where highly localised energy demand is observed through occupant data, the building actuation system must be able to appropriately respond to this localised need. In the case of Dagenham Park School, it was found that some sections of the HVAC system served rooms spanning multiple departments of the school, with varying space uses and occupancy schedules. Without the ability to enact services control choices to the same resolution as occupant monitoring, the energy saving opportunities highlighted by occupant-centric systems cannot be used to their greatest potential.

4.5 Conclusions

The overlying conclusion from the case studies presented in this chapter is a confirmation of some of the issues around occupant data collection that were identified in the review chapters. During the identification of suitable data sources, it was repeatedly found that, while long-term energy datasets were available for a wide range of buildings, the collection of any kind of in-use occupancy data was extremely rare. Where occupant data was collected, it was typically somewhat abstract and would not be appropriate for direct application to building controls without some form of higher-quality verification of the data's reliability to discern actual occupancy.

In a domestic setting, it was found that binary occupant presence can be inferred from single-point motion and CO_2 data, but the reliability of this inferred value is not high. The number of occupants is not possible to infer from this data without some form of ground truth to compare against. In its relationship with energy, water and electrical socket use showed the closest correlation to occupancy rates. However, the poor response of some environmental controls to occupancy was confirmed, with some heating and ventilation systems run without regard to occupant presence. Whether these responses were intended by the building occupants was not possible to

determine with the available data, but would be a highly relevant question to consider when determining the priorities of a more occupancy-responsive control system.

In a large office setting, access control data was used as a proxy for building occupancy, but was not able to provide reliable local occupancy levels due to the lack of coverage on some doors. A potential mismatch between occupants' energy needs and the actual energy use was identified, with occupancy rates showing a stronger weekday to weekend decrease than any of the energy meters. Different office floors showed different correlations to their daily occupancy rate, potentially highlighting more wasteful energy or control behaviours on these floors.

In a school setting, the collection of explicit occupancy data ran into concerns about privacy. Through examination of predefined active school hours, it was found that occupant actions had a significant effect on the energy use relative to expected rates, with higher than expected occupancy outside of contracted hours raising the overall energy use above the accepted threshold for the building's energy contract. Through qualitative interviewing it was seen that the conditioning of some spaces did not adequately account for sudden large changes in occupancy, highlighting the need for slow-response building systems to be able to work pre-emptively where sudden changes in occupant rates can be anticipated.

While each of the occupant data sources in the case studies provided some information, in order to make informed control decisions based on occupancy a more comprehensive source of occupant data is needed. In control application, it is essential that this data source can provide data on both occupants that are expected to be present (such as office staff, who may interact with access systems or carry tagging equipment) and unknown occupants, who can be detected through passive environmental data such as motion and CO_2 level. For this purpose, a multi-sensor model was proposed to combine the benefits of multiple data sources for the predictive system. The following chapter introduces the data sources selected for this study to allow the collection of data on both known individuals and unknown visitors in an office setting.

5 DEVELOPMENT OF SYSTEMS FOR OCCUPANCY DETECTION

5.1 Introduction and Aims

As discussed in Chapter 3, occupancy detection is not a simple concept and requires careful definition of what type of data is needed. In this study, the main aim was to develop a viable way to improve the occupant data provided to building controls, with emphasis on the feasibility of implementing at a wider scale, preferably applicable to existing buildings without intrusive installation requirements. These restraints ruled out some of the more intensive and detailed systems such as RF tagging, which require extensive installation of additional hardware, as well as needing all building occupants to carry tags.

Although some existing studies use highly detailed occupant data, most applications to building controls showed that a localisation to room level is appropriate for most applications. As was seen in the case studies of larger buildings, typical building controls systems are not able to localise environmental controls any closer than to room-level or wider, depending on the system in question. Different building systems favour different occupant data types. Of particular interest are the binary presence of occupants for on/off systems such as lighting, and the number of local occupants in a conditioned space for HVAC applications. Given that local occupancy rates can be easily converted into binary data, it was decided to pursue a system to most effectively provide room-level occupancy rates for this study.

This chapter introduces the testbed used for the major part of this study and the development of sensor hardware/software systems to collect occupant data within this testbed.

5.2 Mark Group House Testbed

5.2.1 Testbed Description

The major data source of occupancy and energy data for the following sections of this project was the Mark Group House, based on the University of Nottingham campus, Nottingham, UK. This building was in use as a small office building for the duration of the project, with single-occupancy and multi-occupancy office rooms alongside shared spaces such as a kitchen and meeting room.

Data was collected in the house using a range of environmental sensors and disaggregated energy meters as described in the following diagrams. The experimental work completed with this data was conducted in two phases, each having a different layout of the occupancy-specific sensors as described below.

Two additional systems were put in place to detect the presence of occupants' personal mobile devices. As these systems can be used to collect personally identifiable data in the case where occupants disclose their device IDs or disclose data collected using beacon-detecting software, an ethics assessment was completed with the University of Nottingham Faculty Research Ethics Committee for engineering to ensure appropriate data collection and storage strategies. An example participation consent form for the named occupants is included in Appendix 10.3. Where personal device IDs were not disclosed, it was not possible to link the data collected to an occupant's real identity – in this case, the detected devices contributed only to an anonymised count of unknown local devices.

5.2.2 Designation of Zones

For the purpose of this study, the building was divided into ten zones of interest. Each of the major spaces in the building that can be expected to be consistently occupied was designated as a single major zone – highlighted in Figure 5-1. Each of these major zones was subject to localised occupancy monitoring. It was decided that restrooms would not be included in the monitored zones due to privacy considerations. All spaces in the building that were not directly monitored, including the restrooms, circulation spaces, storage rooms etc. were designated as a collective final zone, which was not locally monitored, but was subject to the building-wide sensors. A summary of the zones by expected use is shown in Table 5-1.

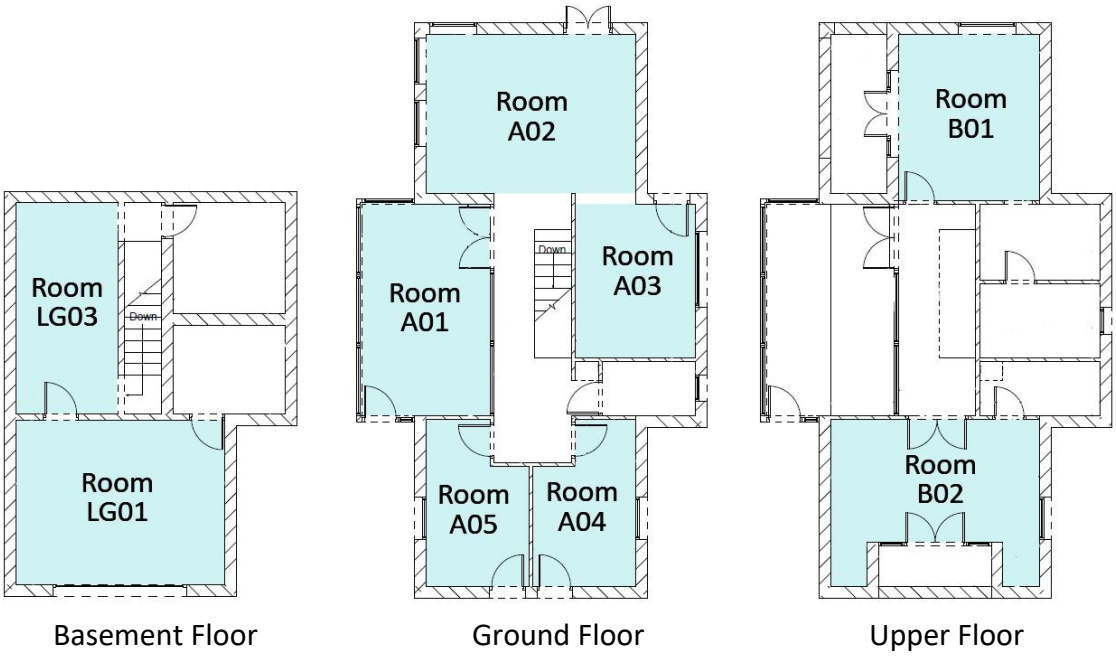


Figure 5-1- Mark Group House Directly Monitored Zones

Table 5-1 - Summary of Mark Group House zone names and uses

Zone Type	Zone Name	Local Monitoring
Single-occupant Office	Room A04	Y
	Room A05	Y
	Room B01	Y
Multi-occupant Office	Room A02	Y
	Room B02	Y
Meeting Room	Room A01	Y
Kitchen	Room A03	Y
Multi-Use Space	Room LG01	Y
	Room LG03	Y
Circulation Spaces etc.	-	N

5.2.3 Phases and Timing of Data Collection

While coverage from most of the installed environmental sensors spanned from late 2013 until early 2017, ongoing issues with the third-party data collection platform and updates to the amount of sensing equipment available meant that the full range of sensors was not consistently available across the entire period of study.

The addition of further sensors to the house part-way through the project split the data collection into two phases:

- Phase 1 – Two motion sensors in total, placed in upper and ground floor corridor areas. Wi-Fi detector placed in upper floor office. Bluebar iBeacons

placed in each monitored zone, but detection software run on only one occupant's device.

- Phase 2 – Eight motion sensors in total, placed in each of the directly monitored zones. Wi-Fi detector placed in upper floor corridor. Kontakt iBeacons placed in each monitored zone, with software run on all regular occupants' devices.

Further details on the changes to the sensor layout made between phases are presented in the following sections. No significant changes were made to the setup of the environmental or energy sensors between phases.

As discussed in more detail in Section 5.2.7 below, it was necessary to collect manually recorded data on the location of building occupants for the work in later chapters. Due to the intensive nature of this manual location recording, this was limited to two week-long periods: one during Phase 1 and one during Phase 2. Table 5-2 summarises the timing of the two test weeks used for the model developed in Chapter 6.

Table 5-2 - Summary of Phase 1 and Phase 2 Data Test Weeks

	Phase 1	**Phase 2**
Manual data collection dates	10/06/2015-16/06/2015	03/02/2017-09/02/2017
No. Days	7	7
Climate Conditions	Spring/Summer Period	Winter Period
No. Regular Occupants Returning Manual Data	9	13

5.2.4 Full Sensor List

Table 5-3 shows a summary of all the sensing devices used in the Mark Group House over the test period, with their locations and their models. Sensing devices have been divided into three categories: environmental sensors, used to contextualise the building's internal environment; occupancy sensors, specifically for measuring occupancy through movement or presence of personal devices; and energy sensors, used to measure the electrical and heating energy use during operation.

Table 5-3 – List of all sensors used in the Mark Group House Installation

Type	Sensor	Location(s)	Sensor Model(s) Used
Environmental	CO_2	All modelled zones	Pressac CO_2, Temperature & Humidity Sensor [185]
	Temperature	All modelled zones	Pressac CO_2, Temperature & Humidity Sensor [185]
	Relative Humidity	All modelled zones	Pressac CO_2, Temperature & Humidity Sensor [185]
	Window	All openable windows	Enocean contact sensor, unknown manufacturer
	Door	All outside-leading doors	Enocean contact sensor, unknown manufacturer
Occupancy	Passive Infra-Red PIR (Phase 1)	B-floor corridor, A-floor corridor	Servodan 41-580 [186]
	Passive Infra-Red PIR (Phase 2)	All modelled zones but LG03	Thermokon SR-MDS [187]
	Wi-Fi Device Detection (Phase 1)	A02, covering whole building	Raspberry Pi B+ [188] with Wireless in Listener Mode
	Wi-Fi Device Detection (Phase 2)	B01, covering whole building	Raspberry Pi B+ [188] with Wireless in Listener Mode
	Bluetooth Beacon (Phase 2)	All modelled zones	Kontakt iBeacon [189]
Energy	MVHR Temperatures	n/a	Temperature probe, unknown manufacturer
	ASHP Energy	n/a	Kamstrup Multical 402 [190]
	Heating Water Energy Use	n/a	Kamstrup Multical 402 [190]
	Hot Water tap Energy Use	n/a	Kamstrup Multical 402 [190]
	Solar Thermal Array Energy	n/a	Kamstrup Multical 402 [190]
	Lighting Circuits	Basement, 2 Ground Floor, Upper Floor, Loft	Current clamp, manufacturer unknown
	Power Circuits	Basement, Ground Floor, Kitchen, Upper Floor, Boiler, ASHP, Solar Thermal, Immersion Heater, Outside Lights, Velux motors	Current clamp, manufacturer unknown
	Photovoltaic Generation	n/a	Current clamp, manufacturer unknown

All sensors but the Bluetooth beacon detection reported data points on value change by default, although for calculation in the models developed in later chapters, the data was converted to a discrete time interval of five minutes. For the count-based sensors such as PIR, this was achieved by summing all data points reported per five-minute period, while for the measurement-based sensors such as temperature this was achieved by taking the mean value of all data points per five minutes.

The choice of a five-minute time resolution was determined as a compromise between the immediacy of response for the proposed control system application, the computational time requirements of running predictive models in real time and a

feasible time resolution at which the model's manually collected training data could be expected. For some control systems such as lighting control, a response would be expected on the order of seconds, however it was not considered feasible for building occupants to accurately manually record their location to the nearest second in this work. Given that existing study suggests that predictive control systems have the greatest benefit for slower response systems such as heating and ventilation, a five-minute interval was considered adequate for this purpose and a realistic interval for manual data collection. For other applications beyond the aims of this work, the processes described in the following chapters could be adapted to a finer-grained time resolution.

5.2.5 Environmental Sensor layouts

5.2.5.1 CO_2/Temperature/Humidity

The distribution of the CO_2/Temperature/Humidity sensors is shown in Figure 5-2. One such sensor was placed in each of the major monitored spaces in the building. The specific placement of the sensors was determined by two factors: strength of light in order to recharge the unit battery, and height to ensure a representative CO_2 measurement. Where possible, sensors were kept at a consistent height across all zones.

Figure 5-2 - CO2/Temperature/Humidity Sensor Placement in the Mark Group House

5.2.5.2 Door/window

The placement of door/window sensors was constrained to openings to the external atmosphere, as demonstrated in Figure 5-3. This was intended to allow for analysis of air flows relevant to heating/ventilation energy usage.

Figure 5-3 – Door/Window Sensor Placement in the Mark Group House

5.2.6 Occupancy-specific Sensors

5.2.6.1 PIR Motion Sensors

The layout of the PIR motion sensors was varied between data collection phases 1 and 2. Figure 5-4 shows the layout used in Phase 1,where the two available motion sensors were placed in the ground and upper floor circulation spaces for maximum coverage across the used spaces of the building with limited sensors. For Phase 2, Motion/PIR sensors were placed in each major monitored space, shown in Figure 5-5. However, the unit intended for Room LG03 was found to be non-functional, and so has been omitted from this study. The PIR sensors were placed with the intention to maximise coverage of the room, while minimising interference from neighbouring spaces. The range of each sensor was approximated using the manufacturer's data at a given ceiling height, and is indicated in the layout diagrams. As shown in the figures, there may be some issue with movement outside some windows causing false positive readings, however where possible this has been kept to a minimum.

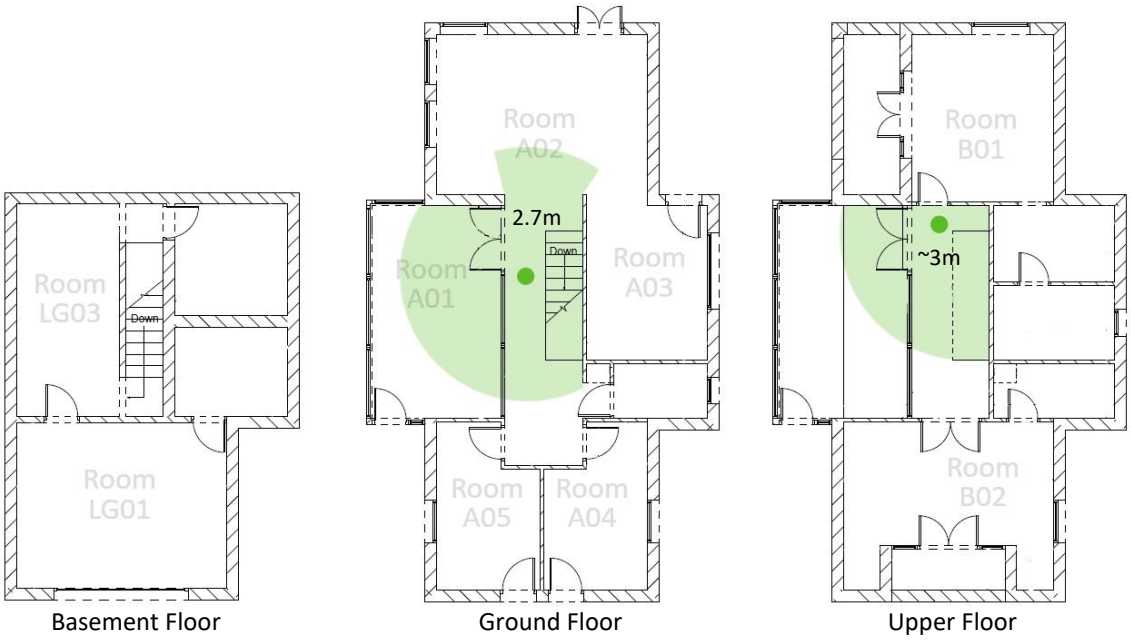


Figure 5-4 - PIR Sensor Placement and Range in the Mark Group House, Phase 1 Layout

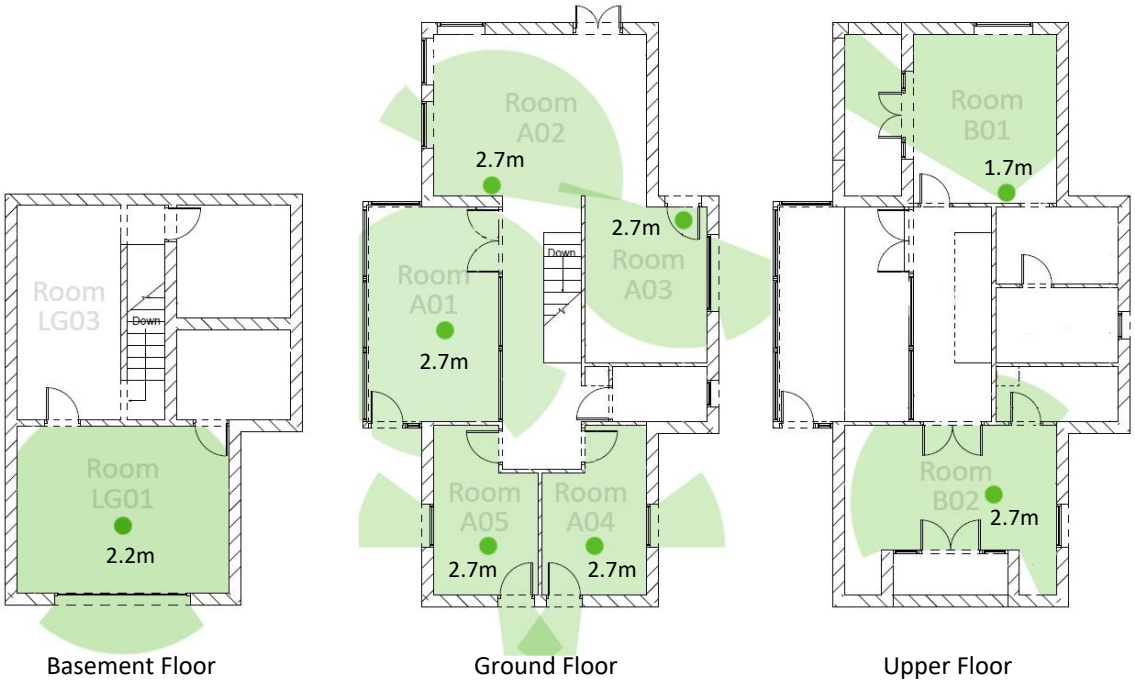


Figure 5-5 - PIR Sensor Placement and Range in the Mark Group House, Phase 2 Layout

5.2.6.2 Wi-Fi Detection and Bluetooth iBeacons

Bluetooth low-energy (BTLE) iBeacons were, once again, placed in each monitored zone as shown in Figure 5-6. Where possible, beacons were placed on an outside wall, such that each beacon was as far as possible from other beacons. In particular, beacons on the ground floor were placed low on the wall where possible, while upper

floor beacons were placed high on the wall. This was intended to minimise areas where two beacons can be detected with equal signal strength. During Phase 1 of data collection, Bluebar iBeacons were placed in each zone, but were used only for testing of the beacon location software on one occupant's personal device. For Phase 2, the more consistent Kontakt iBeacon hardware was installed in place of the original beacons, and was used for data collection across all occupants.

Wi-Fi detection used a device placed at a single point in the building to listen for Wi-Fi devices throughout the whole house. The upper floor location was chosen to minimise the signal strength received from devices outside of the building at ground level – the closest of which should be detected through both floor and wall material, rather than wall material only if the detector was placed on the ground floor. Further details on the placement of the Wi-Fi detection device can be found in section 5.4.2.

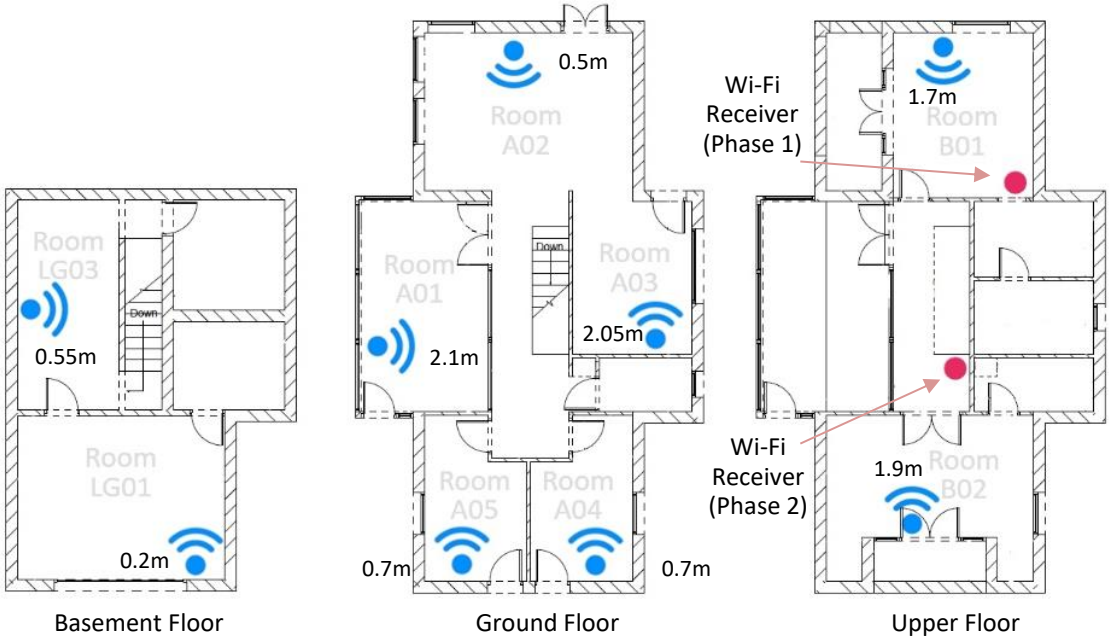


Figure 5-6 - iBeacon & Wi-Fi Detector Placement in the Mark Group House

5.2.7 Manually Recorded Occupancy Data

In later sections of this study, it was necessary to use manually recorded location data from building occupants in order to assess the effectiveness of the Wi-Fi and Bluetooth based personal device detection and to train models used to combine sensor data into an estimate of the local number of occupants. A one-week period of manual location data collection from all building occupants was conducted for each of the two

sensor layout phases. Appendix 10.4 shows an example of the datasheet used to collect this information.

Data was collected to the nearest 5 minutes, which was considered the best compromise between obtaining information precise enough to match the sensor data resolution, while keeping the amount of work required by the participants to a realistic level. Initial informal surveying of occupants had suggested that a data collection process that took too much effort would discourage some occupants from participating.

As the responsibility for maintaining accurate results from this manual data collection was reliant on each individual occupant, it was expected that there would be some variation in response from person to person. In the first phase of testing, two occupants did not complete the assessment at all and so their respective zones were omitted from the phase 1 model. In the second phase of testing, occupants were issued with questions on their experience with the data collection process, with the received responses shown in Table 5-4. It can be seen that the majority of occupants felt that they were able to provide location data to the required 5-minute accuracy, but it should be noted that some respondents felt their data was less accurate than this. In comments from some respondents, it was said that longer periods of occupancy in the same space were likely more accurate than short stays, some of which were forgotten before they were logged on the datasheet. This is an important factor to consider in the model training sections, as inaccuracies in the training data could potentially create an inaccurate model. It is assumed that, with enough positive responses in similar situations, a small subset of inaccurate data points in the training set should be treated as outliers in most trained model types, without a significant effect on the overall model performance. However, in cases where few examples of a particular situation exist, these inaccurate training data points may present issues for the accuracy of the model.

Table 5-4 - Occupant Responses to Manual Location Data Collection

Occupant Type	Occupant Main Zone	Datasheet completed During or After Events	Do you feel you achieved 5-minute level accuracy?
Student	A02	During	Yes
Student	A02	During	Yes
Researcher	LG01	During	Yes
Researcher	LG01	During, mostly	Yes
Researcher	LG01	After	No
Researcher	A02	During	Yes
Teaching Staff	A04	After	No
Teaching Staff	B01	During	No
Teaching Staff	A05	During	Yes

5.3 Selection of Appropriate Data Collection Methods

There is a large variety of methods to detect occupancy available, as discussed in the literature review. Each data source has characteristics that determine how appropriate the sensor type is for a given situation – whether occupancy needs to be detected, counted or classified, and what type of building space is being monitored. The data collection for this project is centred on a small office space, but has the added complexity of being a demonstration building on a university campus, meaning that the space is often open to groups of temporary visitors for academic meetings, building tours etc.

As this study is investigative, it was planned to collect as wide a range of occupancy data as possible in order to compare the value of data from each source and test how much information can be gained in the combination of data sources. As described in section 5.2, environmental sensors were placed in each monitored zone of the building. Additional detection for the personal devices of building occupants was identified as a valuable addition to the dataset, as this can be used for localisation of individual occupants beyond the capability of environmental sensors alone, but does not require extensive use of additional sensing hardware.

The detection of personal devices typically uses some form of wireless signal: telephone network, Wi-Fi or Bluetooth. Previous studies have shown the viability of using Wi-Fi connections from personal devices as a proxy for building occupancy levels [52]. One of the shortfalls identified in the Wi-Fi network based solutions shown in this study is the ability to detect the devices of people previously unknown to the system. Indeed, in many popular detection methods, devices not running specific software or not connected to the local Wi-Fi network are not able to be

counted. It is also known that some commercial systems exist to estimate footfall in retail applications [191]. These systems work by detecting the MAC address – a unique device identifier used for internet connection. Such systems are typically used as a way to analyse broad trends in occupancy over time, rather than accurate localisation in a small space.

It was also decided that a more localised room-level system would be implemented alongside the Wi-Fi traffic detection, which, sensing from a single point, would not be able to locate occupants at a room level without further hardware. Sensing full location in 3-dimensional space is highly complex and costly to gather data and to analyse [80] and is not a viable option for most commercial applications. The use of localised Bluetooth Low-Energy (BTLE) beacons and mobile phones was designated as a good solution for an office application, as:

- Office occupants can be assumed to have a mobile phone present with them in the building for the majority of the time.
- Privacy issues related to the collection of personally identifiable data are somewhat negated, as the system collecting data is the individual's mobile phone, meaning that any data stored can be personally reviewed and data is only collected from occupants who have opted to use the location software.
- Bluetooth Low Energy allows for a more viable use of occupants' mobile phones with minimal disruption to battery life from frequent scanning in comparison to earlier versions of the Bluetooth standard.
- A minimal amount of extra hardware needs to be installed – each zone only requires one small wireless beacon.

5.4 Development of Data Collection Methods

5.4.1 Raspberry Pi Wi-Fi Detector Setup

In this study, a smaller-scale, low-cost solution was explored using the Raspberry Pi – a small, inexpensive, Linux-based computer board. Using a Wi-Fi monitor-mode compatible USB Wi-Fi adapter, it is possible to scan for the presence of Wi-Fi traffic in the surrounding area. This has the benefit of detecting occupants who are not previously known to the system and is appropriate for a small office space. In a larger space, the degradation of Wi-Fi signal strength over distance and through building material would mean that multiple devices would be needed to cover the area.

The detector was set up using a standard Raspberry Pi B+ board and USB Wi-Fi adapter compatible with monitor mode – in this case, an adapter using the Ralink RT5370 chipset. A list of compatible chipsets was found in the support information of an open-source software package utilising Wi-Fi monitoring [192].

Using open-source software to interface with the Wi-Fi adapter [193], a script was written to collect the MAC address, signal strength and connection type of devices broadcasting Wi-Fi probe signals within range of the detector. The script was based upon an open source solution using the same monitoring software [194].

The raw data from the Wi-Fi detector required pre-processing before meaningful information on occupancy could be extracted, including elimination of static devices such as Wi-Fi routers and identification of people passing outside the building.

5.4.2 Raspberry Pi Wi-Fi Detection Testing

5.4.2.1 Analysis of a Small Office Setting

An analysis of the Wi-Fi detection data was conducted on a week with manually labelled occupancy for the building – detail on how this data was acquired can be found in section 5.2.7. The unprocessed data from the Wi-Fi detector shows a high density of devices detected, with 0.7 million detection events from 888 unique device IDs detected over a 7-day period. Over 78% of such events are from 5 devices, as illustrated in Figure 5-7. These are likely static background devices - defined as devices that have visible Wi-Fi traffic, but are in a fixed, static location in the building and thus are not related to occupancy levels. Examples include Wi-Fi network routers, Wi-Fi enabled display PCs etc. Other signals received at low-strength for a short time can be assumed to be from people passing by outside the building. It can also be expected that some building occupants carry multiple Wi-Fi enabled devices, and some carry none.

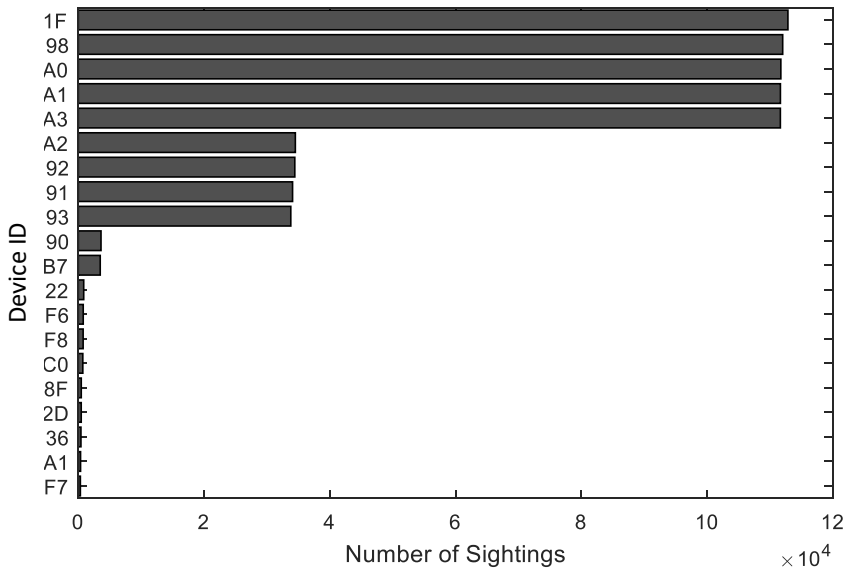


Figure 5-7 - Number of Sightings of the 20 Most Frequently Detected Device IDs, Phase 1

Static devices can be omitted by identifying when the building is empty of occupants, and creating a 'blacklist' of MAC addresses to be ignored from the IDs still found during this time. Two methods were tested to identify when the building was empty: manual designation of times believed to be sufficiently outside of office hours to guarantee zero occupancy, and automated designation of zero occupancy based on environmental sensor measurements. The first method is limited to a short period of time when the working patterns of occupants was known by survey and may lose accuracy over time as devices in the house are changed, while the automated method has the potential to falsely blacklist genuine occupant devices if zero occupancy is incorrectly assumed.

Manual identification of static devices yielded a list of 11 MAC addresses that were present during the hours of 00:00 to 04:00 during the test period. Automated identification of static devices was based on the environmental readings of CO_2, PIR and cold water use. These three measures were found to be as the best indicators of occupancy in the domestic case study covered in section 4.2. Non-occupied periods were assumed when the following conditions were met:

- Zero motion count from all PIR sensors in the last 60 minutes.
- CO2 level below 700ppm in all zones (value determined by mean value when known unoccupied during test weeks).
- Zero cold water use recorded in the last 60 minutes.

The assumed non-occupied periods are shown against the verified non-occupied periods in Figure 5-8. It can be seen that the automatic detection is intentionally more prone to false negatives than false positives, as omitting some data from a non-occupied period is less of a problem than blacklisting IDs from an occupied period. For the test week data, 24 devices were identified. This included two MAC addresses known to belong to occupants of the building shown as crosses in Figure 5-8. Both false positive detections occur at the start of a day, implying that there is some lag in the detection of occupancy using the PIR, CO_2 and water sensor combination with the equipment arranged as in the Phase 1 test period. To counter this problem, the last 15 minutes of each automatically detected non-occupied period was ignored.

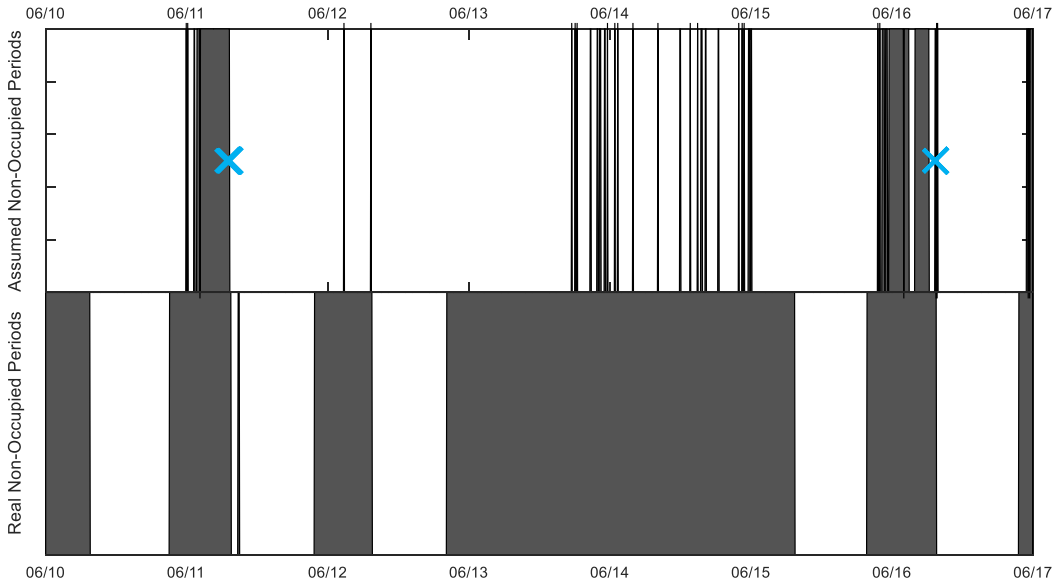


Figure 5-8 Real vs Assumed Periods of Zero Occupancy

A total of 9 devices were on the blacklist of both methods for Phase 1 of testing. These IDs were the top 9 most detected MAC addresses, representing 97% of all detection events. Omitting this list of devices from Phase 1 data processing both removes factors irrelevant to occupancy rates and significantly reduces the computational load for calculations on the Wi-Fi data. During the testing for Phase 2, it was found that the most commonly seen static devices had changed (Figure 5-9). In a working building, it is reasonable to expect that static devices in the building will change over time. For this reason, the detection model developed in Chapter 6 was set up to automatically identify static devices in the given Wi-Fi data at the start of any training, using the methods described above. In practice, an occupant data model run

over long periods without retraining would need the blacklist of static devices to be periodically updated using historic sensor data.

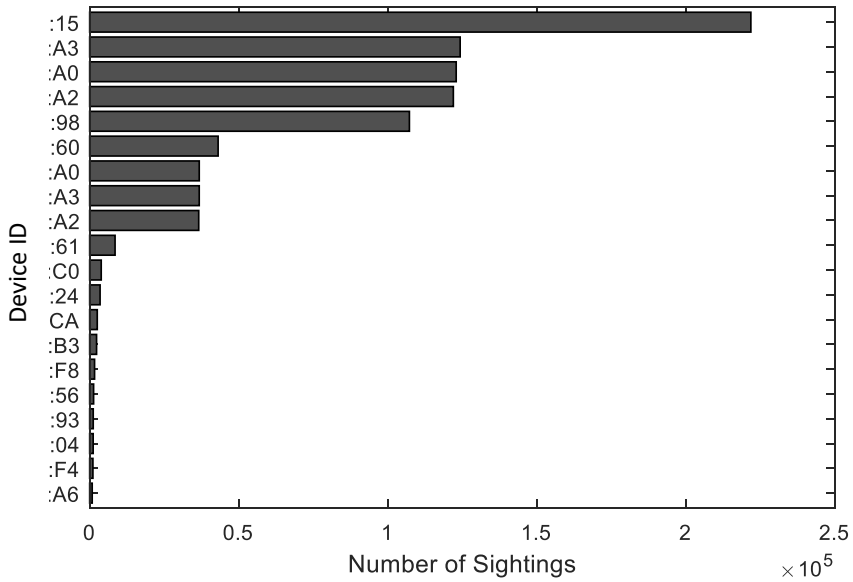


Figure 5-9 - Number of Sightings of the 20 Most Frequently Detected Device IDs, Phase 2

Once the static devices were omitted, the Wi-Fi data had to be processed. Each device detection represents a single point in time; in order to translate a series of distinct detections to continuous states of occupancy/non-occupancy, a reasonable gap between detections must be assumed. In practice, this takes the form of a period after each detection where the device is assumed to still be present, but silent. The average length of time between detections was investigated over the test week in order to find an appropriate length for this assumed period. Figure 5-10 shows a histogram of the duration between detections for all devices over the entire test period. It can be seen that the majority of detections occur with a frequency below 20 seconds, however, this data is skewed by the prevalence of signals from background devices as discussed above. A more accurate frequency for devices of interest can be found by measuring the time between detections for devices belonging to known occupants of the building during only hours these occupants are verified to be present. Figure 5-11 shows the results of this analysis. It can be seen that the majority of detections occur with an average frequency of 10 minutes or less: with 80% of detections occurring within 5.3 minutes of each other, and 90% known detections within 10.1 minutes. This suggests that assuming an occupant is present for 5-10 minutes after each detection is a reasonable threshold to estimate continuous occupancy. The threshold value for assuming continuous occupancy was set at 10 minutes.

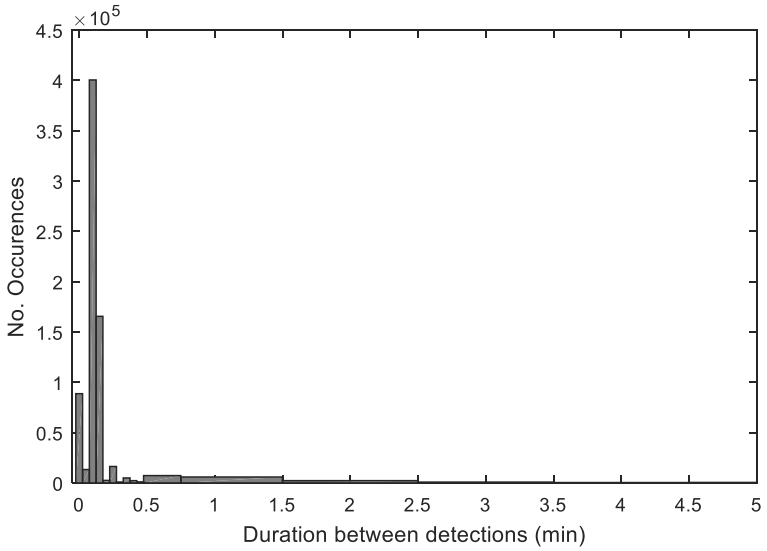


Figure 5-10 - Length of time between detections for all devices

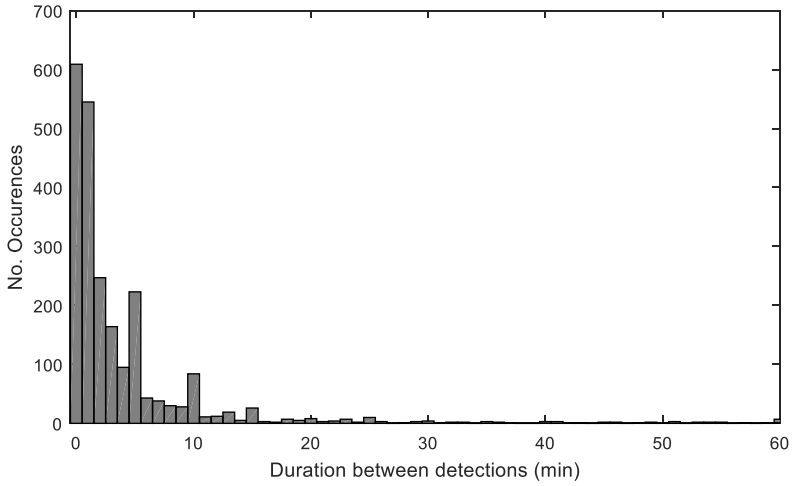


Figure 5-11 - Duration Between Detections for Occupant Devices Known to be Consistently Present

Figure 5-12 shows the results of Wi-Fi detection over a single day in the Mark Group House, showing only regular occupants of the office whose MAC address was manually identified.

Table 5-5 Summary of Individual Wi-Fi Detection Frequency while Present for the Test Week

Occupant	Mean Gap Between Detections (min)	90th Percentile Gap Between Detections (min)	Device OS
Person E	31.5	108.0	Android
Person D	13.4	36.1	iOS
Person A	11.7	30.0	Android
Person I	7.4	2.4	iOS
Person L	5.3	8.0	Android
Person C	4.6	10.1	Android
Person F	4.5	7.7	iOS
Person B	3.8	8.5	iOS

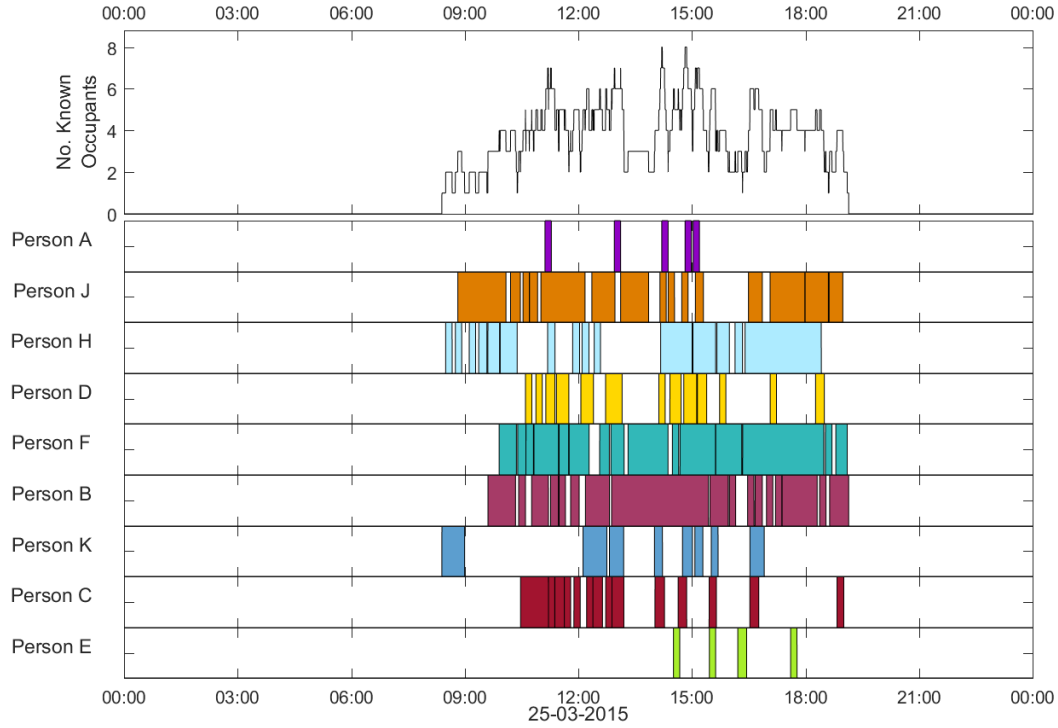


Figure 5-12 - Example day of Wi-Fi based detection - known office occupants only

It can be seen that, despite the threshold optimised to cover 90% of the average detection frequency, some devices are detected consistently throughout the day, while others are more sporadic. This is likely due to:

- Operating system specifics – those with iOS based devices showed a general tendency to be more consistently visible than Android, as shown in Table 5-5. Android devices in this study were more likely to leave large gaps between Wi-Fi activity. Default settings on data use and updates specific to the operating system may have influence on the visibility of devices.
- Setup of the device in question – it is not a guarantee that occupants will have their device's Wi-Fi enabled at all times.
- Software/apps running background processes and updates – those with a larger number of apps using internet connection (email checking, automatically obtaining weather forecasting/news etc.) will experience greater background levels of Wi-Fi traffic than those without.
- Personal device usage patterns – it is expected that some occupants will use their devices for internet-based tasks more often throughout the day than others, allowing more visible traffic.

While the varying coverage of different individuals is a potential hit to the reliability of the passive Wi-Fi listening method, it was not considered feasible to try to calculate more personalised thresholds for assuming continuous occupancy. One significant reason for this is that longer assumed occupancy could begin to introduce false-positives to the data, where an individual is recorded as present when they are actually absent. As the data currently stands, it should only have false negative errors for the vast majority of cases, as it is physically impossible to detect signals from a device that is not present. As part of a wider system of inference, this bias against false positives provides some information in itself: if a user's device is detected, the user is most likely present, but if not detected, other data may be required to gain more certainty. A general threshold of 10 minutes (or two timesteps of a 5-minute resolution system) over all devices was considered the best compromise between continuous detection and avoiding false positives from the Wi-Fi detection.

5.4.2.2 Signal Strength Received in a Small Office Setting

During the initial testing of the Mark Group House, a correlation was found between the total number of devices detected over a threshold signal strength and the total number of people in the building at that time, including guests previously unknown to the system. However, this initial test did not include any significant number of occupants in the basement areas of the building. A more comprehensive test of the signal strength achieved was conducted at the extreme points in the building, to test whether the previously found threshold is appropriate for all zones equally.

A series of spot tests of signal strength were conducted using a single mobile device with known MAC address. The Wi-Fi functionality of this device was switched off and on several times, prompting an outgoing signal at reliable intervals that could then be identified from the Raspberry Pi data. The Raspberry Pi data collection device was placed in an office on the upper floor of the building, as shown in the diagrams in section 5.2.

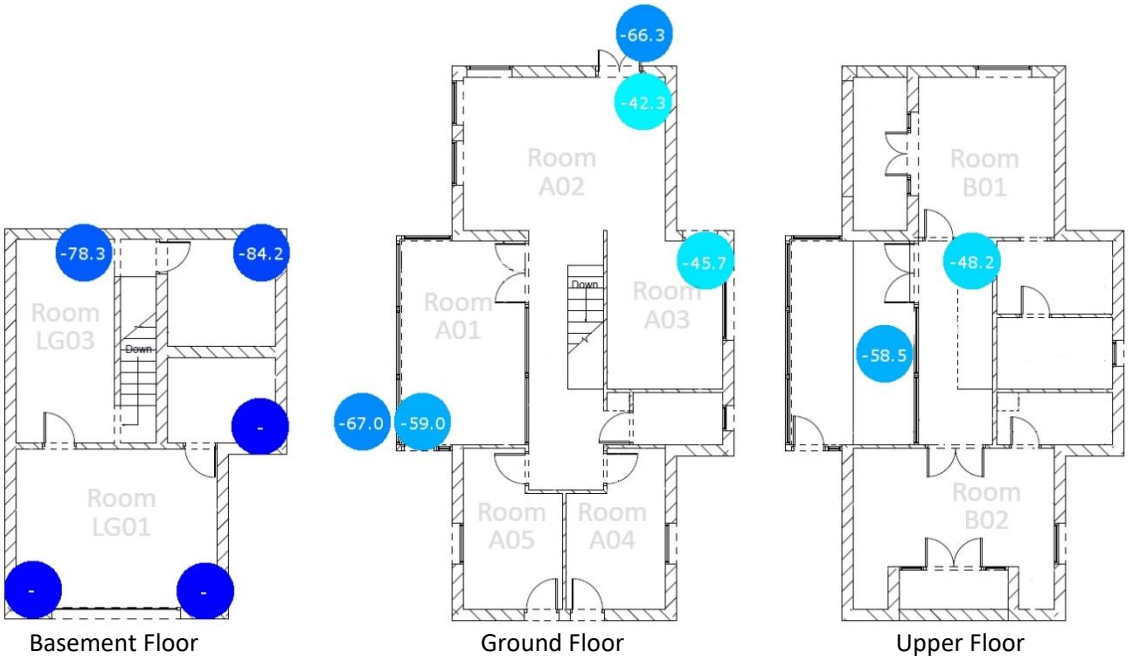


Figure 5-13 - Results of Wi-Fi Detector Signal Strength Spot Test in the Mark Group House

The averaged results of the spot tests are shown on the building plans in Figure 5-13. It can be seen that a signal strength threshold of -65dbm would be a reasonable cutting off point for occupants on the ground or upper floor, as all measurements from within these spaces are stronger than this value. Measurements from outside the building walls all fall below this value. However, the measurements taken in the basement zones show signal strengths significantly below the previous threshold, with the most extreme locations not registering any detected signals at all. It should also be noted that these spot tests were conducted with a single device only: it can be expected that different devices may output Wi-Fi signals of varying strengths, meaning that the -65dbm threshold would not be universal for all occupants.

5.4.2.3 Analysis of a Large Occupied Space

The Mark Group House, as a small office in a single building, represents an uncommon case where the area of monitored is physically independent of any other occupied spaces. In other applications, it might be expected that adjacent spaces, external footpaths etc. will produce interference in the local device detection.

It was decided to test the Wi-Fi detection in an internal space on the university campus during a lecture period. Figure 5-14 shows the room highlighted in blue. The space is surrounded by 5 other lecturing spaces, all of which were in use during the

test period. The test was run from 10:55 to 13:20, covering a 2-hour lecture scheduled from 11:00 to 13:00.

The placement of the detection hardware is shown on Figure 5-14. Due to the need to site the hardware next to a power supply, it could not be placed centrally.

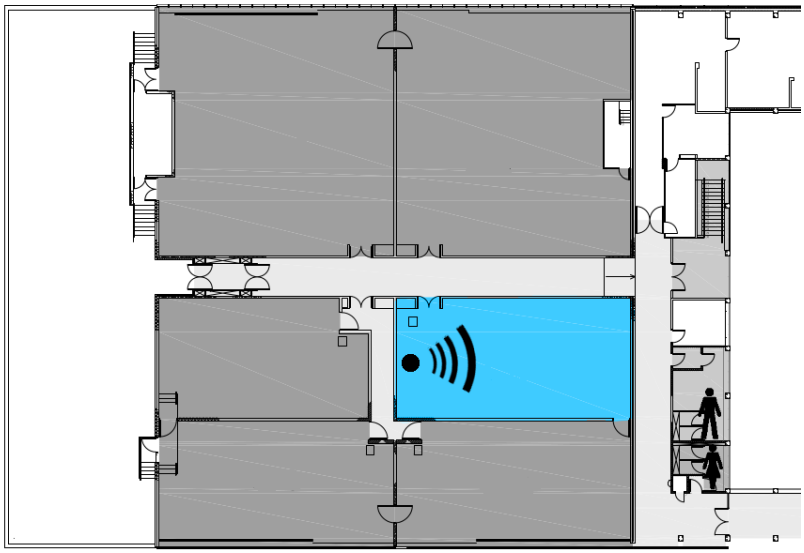


Figure 5-14 - Floor Plan of Test Space

Background devices are defined as devices that have visible Wi-Fi traffic, but are in a fixed, static location in the building and thus are not related to occupancy levels. Examples include Wi-Fi network routers, Wi-Fi enabled display PCs etc. These devices were identified during a secondary testing period, taken 18:30-19:00 on the day following the initial sample. During this time, the lecture spaces were unoccupied. In an ideal situation, the building would be fully vacated of occupants for this test. However, due to restrictions on the opening times of lecture buildings, the test could not be conducted with a guarantee of zero occupant devices in range beyond the lecture halls. It was assumed, however, that any devices detected more than once during this period were background devices and so were omitted from the following analysis.

Figure 5-15 shows the majority of detected devices broadcast with a mean frequency of under 5 minutes. The average length of time between detections for 90% of the devices was 10.4 minutes. This suggests that the optimal length of time a device can be presumed present after an instantaneous detection is between 5 and 10 minutes: in the following graphs, a value of 5 minutes was used.

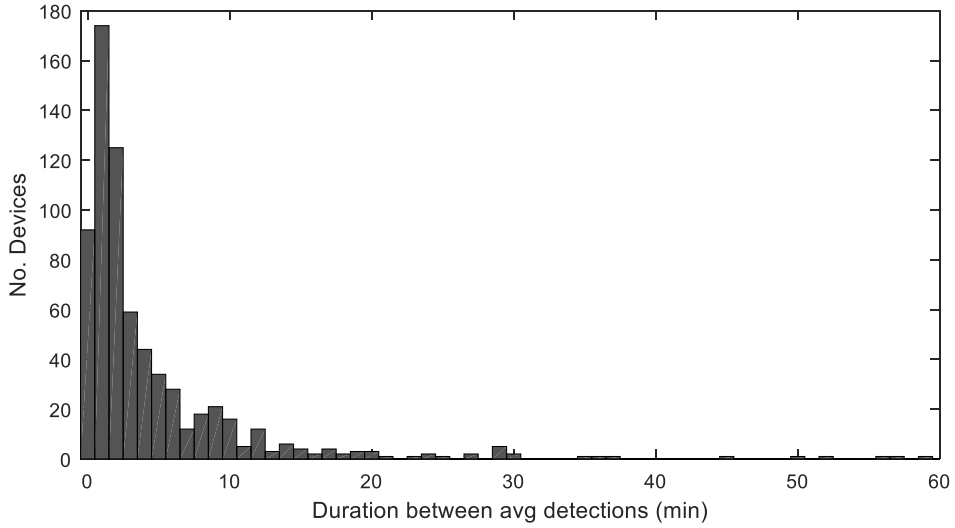


Figure 5-15 - Histogram of average length of time between detections, per device

Figure 5-16 shows the total number of non-background devices detected over the period. It can be seen that the total number of devices far exceeds the number of people in the space. This confirms the effects of interference from other adjacent spaces.

However, the pattern of the data over time can be mapped to observed events in the tested space and adjacent spaces. This means that, while the magnitude of the number of devices detected is not a perfect measure of occupancy, there is information to be gained from analysis over time. Seven events have been identified in the data below:

1. Peak as people gather in the corridor space for the next lecture period
2. Drop as previous lectures empty from the lecture rooms
3. Relatively stable period, lectures continue
4. Peak/drop as lectures ending at 12:00 switch or leave
5. Lower stable period, indicating an adjacent room is no longer in use
6. Peak as people gather in the corridor space for the next lecture period
7. Staggered drop as lectures empty

The higher signal strength detections (-60dbm and above) show peaks at the start and end of the lecture period. However, there is also an unexplained series of peaks/troughs in device detections between these two points in time. This could be due to several factors, such as a lecturer periodically walking in front of the sensor, or

potentially some synchronisation of data use cycles across a large number of devices. Without further data, however, the source of the peaks cannot be determined.

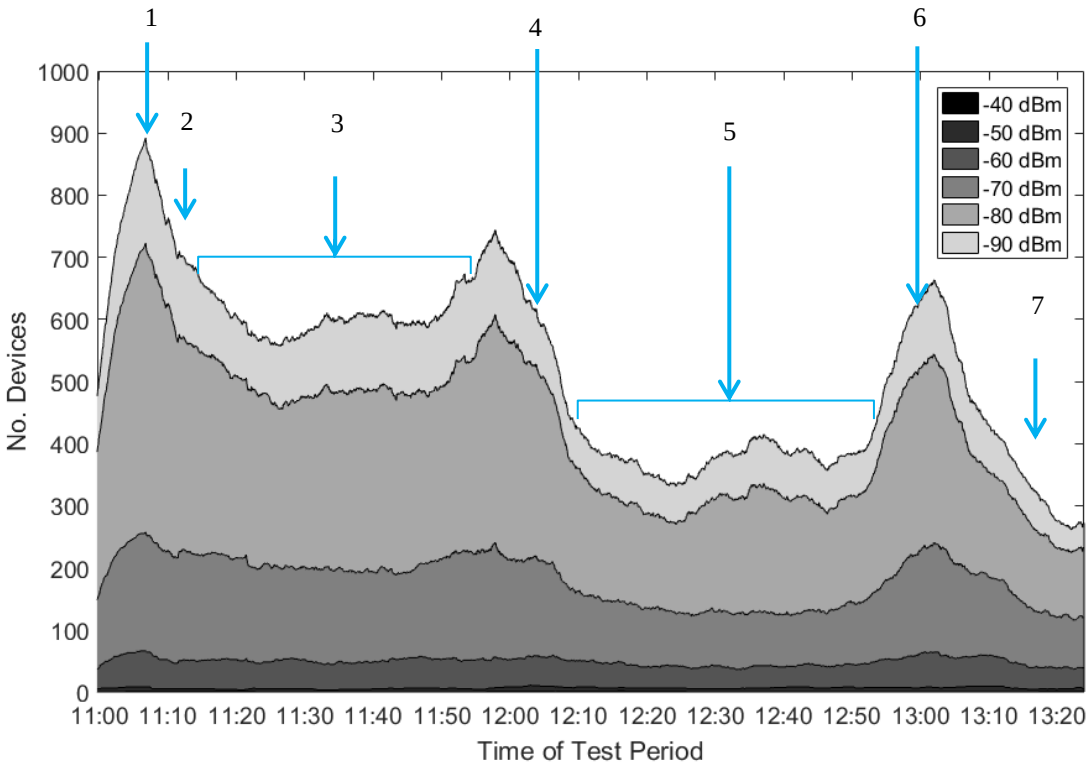


Figure 5-16 - All Devices Captured During the Test Period

The total duration of stay for each device was also calculated – this being the length of time between the first and last detection made. Figure 5-17 shows the results of this analysis for devices of signal strength -70dBm or higher, with background devices and devices seen only once during the period omitted. It can be seen that a high number of devices are detected for less than 10 minutes – likely these devices were held by people passing the lecture hall or further away in the building. A sharp peak at 50 minutes shows the devices of people in neighbouring lecture halls who arrived and left during the first hour of the test. A smaller peak around 120 minutes shows a significant number of people who were present over a two hour period – likely occupants of the tested lecture hall during the two-hour lecture.

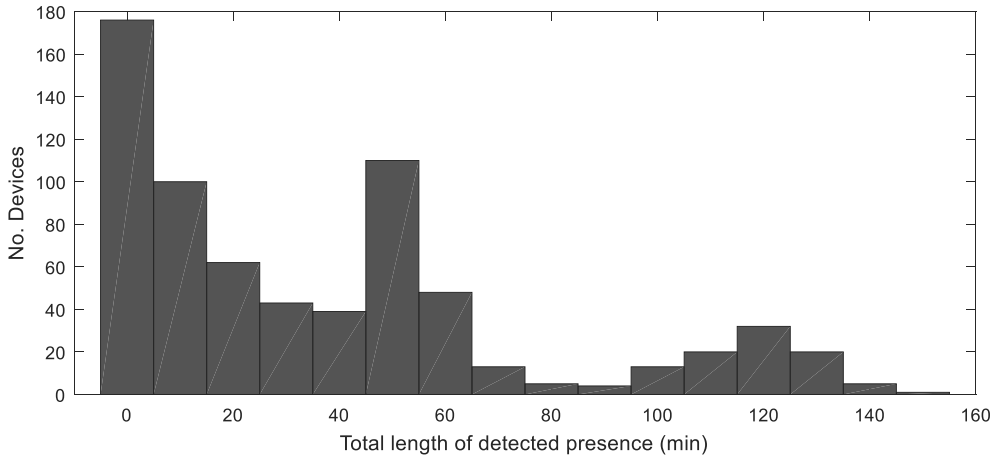


Figure 5-17 - Histogram of total detected presence time

5.4.2.4 Signal Strength Received in a Large Occupied Space

In order to better understand the signal strength attenuation by distance and building material around a monitored sub-space as part of a larger building, a series of spot tests were conducted during the out-of-hours test, in a similar manner to the spot testing in the Mark Group House. As Figure 5-18 shows, signal strength was generally attenuated more by walls than distance, as expected from the results in the Mark Group House. However, an occupant in an adjacent room close to the sensor could have higher signal strength than someone in the back of the test room. This situation suggests that, if wishing to measure the occupancy of a single room with Wi-Fi detection, care should be taken to locate the sensor as centrally as possible in the room, although this may not always be possible. Another potential solution is to use multiple Wi-Fi sensors and attempt to triangulate signals detected.

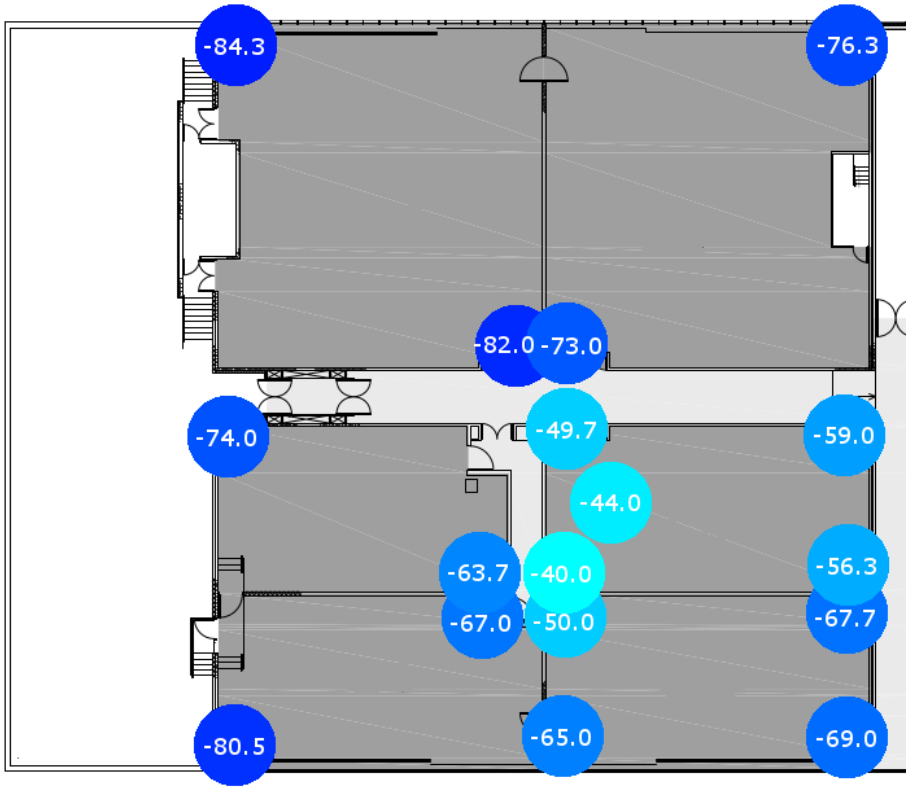


Figure 5-18 - Signal Strength Spot Test Results in the Lecture Hall

5.4.2.5 Discussion

Initial analysis of the results suggests that Wi-Fi probe signal detection can provide valuable information on building occupancy patterns and is able to provide some measure of non-regular occupants who are not connected to the local Wi-Fi networks. However, due to the infrequent detection of some devices, it cannot be used alone as a consistently accurate occupancy measure and thus must be used as a supplement to other sensing systems.

The system also detects any Wi-Fi connected devices including local routers and other permanently placed devices. In public spaces that are closed at night, these can be identified during periods with no occupancy. The signal strength received can also be used to filter out false positive results from personal devices being carried outside of the monitored area, but this is dependent on the location of the sensor and makeup of the building.

In spaces with a large population, the trend in number of devices seen gives potentially more information than the actual number of devices seen, which was found to be significantly greater than the local number of people.

5.4.3 iBeacon Hardware Setup

In theory, all iBeacon hardware should have similar performance under the iBeacon protocol. However in practice, this was found to not be the case. Initial tests were conducted with Bluebar USB iBeacons [195], supplied with power from mains electricity adaptors. This hardware had the benefit of not relying on batteries for power, meaning that the longevity of the signal output should not be an issue over a long time period. However, it was found that the performance from beacon to beacon was highly variable, with some beacons failing to send out a consistent signal after several weeks of use.

As an alternative hardware, Kontakt [189] beacons were then selected, as they were these were found to be the most consistent between different beacons in reviews of several beacon types [196], and have a reasonably long battery life compared to other beacon brands [197]. In an application where it is likely that multiple beacons can be detected at once, it is important that the signal strength received from each beacon is consistent with its distance from the user.

Both beacon types were sampled in the Mark Group House for a period of 40 minutes using an Android-based beacon scanner. It was found that all Kontakt beacons tested emitted signals at a more consistent interval, shown in Figure 5-19a, while the Bluebar beacons were inconsistent, with some beacons showing a profile similar to the Kontakt, some showing larger gaps between signal emissions as in Figure 5-19b, and some not emitting signals. It was therefore chosen to use the Kontakt beacons only for further testing in the Mark Group House.

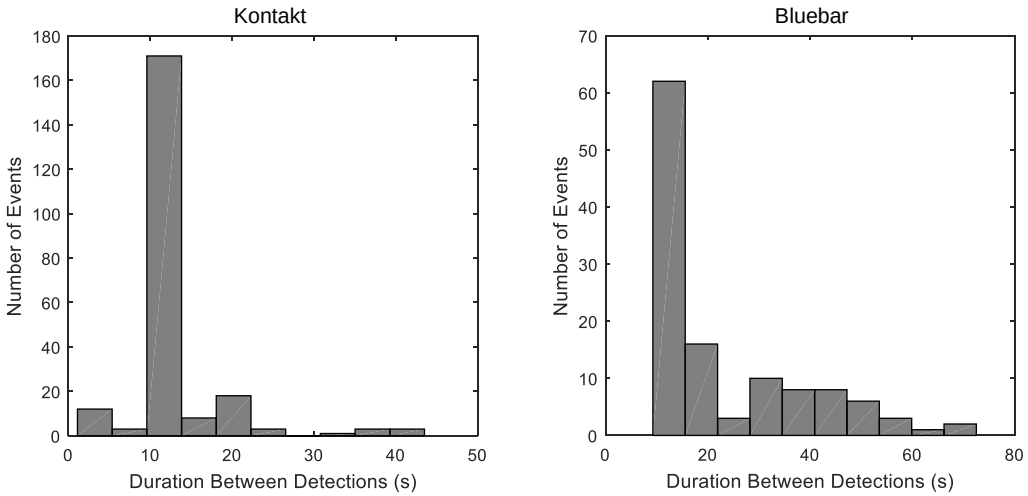


Figure 5-19 Sample Detection Rate for a) Kontakt and b) Bluebar iBeacons

5.4.4 iBeacon Software Testing - iOS

A survey of occupants in the test building revealed that all regular occupants owned either an iOS or Android based device with Bluetooth. For iOS devices, the app 'Geofency' was found [198], which covered the majority of requirements and so was selected for iOS users.

When the unique identifiers of an iBeacon are entered into a user's Geofency app and the device Bluetooth enabled, the app will log an entry/exit time for when the user is within detection range of the iBeacon. This functionality can also be enabled for GPS locations, although this was not required for the study. It should be noted that each location in the app is treated independently of the others, meaning that there is no logic applied when multiple iBeacons are in range. The user is simply logged as present in multiple locations. For the purpose of locating an individual within the Mark Group House, this means that the combination of zones visible to the device will likely be a better indication of location than just recorded presence in the zone of interest.

Figure 5-20 shows a sample of data collected from Geofency during the first Phase of experimentation: with one BlueBar beacon per zone. It can be seen that there is some agreement between the Wi-Fi detection and the bluetooth, but that the user is logged into two zones simultaneously for most of the occupied period. The user also becomes 'stuck' in one zone overnight, despite the occupant not being present: this appears to be an artifact from the Geofency application caused by some incorrect sign out of the zone. The 'stuck' zone is corrected only when this beacon is encountered again the next day. The overnight errors are easily indentified, however, it should be noted that these errors could also occur for shorter periods during the day. This means that, although it was previously expected that the rate of false positives from this method would be low to zero, 'stuck' beacons could introduce false positive readings that are not easily identified.

Over the period of one week, the data collected from Geofency and Wi-Fi detection were compared for a single user. As an estimation of binary occupant presence (presence assumed when the user was logged into any zone), the methods agreed for 89.1% of the text week. Given the long periods of absence overnight and at weekends, this value may appear to be optimistic. As a measure of agreement during occupied

periods only, it was assumed that the user was present any time at least one method reported presence. During ‘occupied’ times only, the amount of time where both methods reported presence was 48.0%. This much reduced number may be due in part to the false positive overnight values in Geofency.

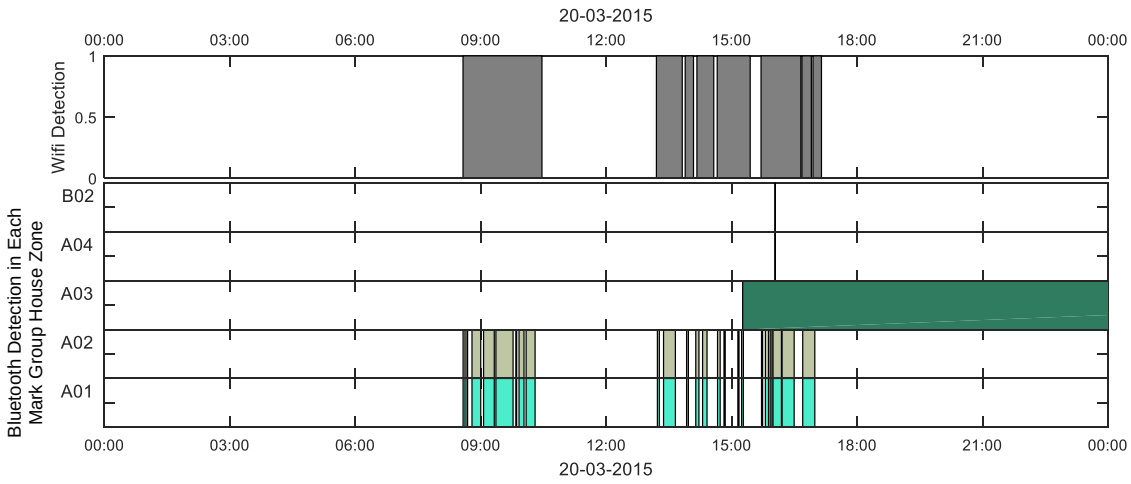


Figure 5-20 - Geofency Results vs Wi-Fi Detection - Phase 1 Bluebar iBeacons

In the second phase of testing, the beacon hardware was switched to Kontakt iBeacons, and the Geofency app was implemented on the devices of 6 building occupants. During this time, the occupants recorded their location manually for reference. The reliability of this manual recording is discussed in section 5.2.7. It can be seen that Geofency showed a level of accuracy to the self-reported presence of occupants that was consistently above 90%. Wi-Fi detection was more variable, as the consistency of detection depends on device use patterns, discussed above. For example, occupant D can be detected consistently with the Geofency data, but is only detected sparsely with the Wi-Fi detector (Figure 5-21). It can be seen in Table 5-7 that Wi-Fi detection was generally more prone to false negatives, while the Geofency detection was more prone to false positives, likely due to the ‘stuck’ zone problem that was identified in Phase 1.

It should be noted that all occupants tested in Phase 2 found that Geofency logged their device into multiple zones consistently whenever they were present in the building. Therefore, the accuracy of detected zone cannot be easily quantified.

Table 5-6 - Summary of Geofency vs Wi-Fi Agreement per Occupant

Occupant	Wi-Fi-Geofency Agreement	Wi-Fi-Geofency Agreement while assumed occupied	Geofency Accuracy with Self-reported Occupancy	Wi-Fi Accuracy with Self-reported Occupancy
Person B	96.5%	42.1%	98.0%	96.7%
Person D	78.7%	12.4%	99.0%	79.3%
Person F	98.1%	31.6%	99.6%	97.8%
Person J	90.4%	40.3%	90.1%	90.4%
Person K	92.0%	63.4%	97.7%	91.8%
Person P	94.6%	52.2%	99.2%	94.1%
Average	**93.1%**	**49.7%**	**97.2%**	**93.1%**

Table 5-7 - Summary of False Negative/Positive Rates for Geofency and Wi-Fi Detection

Occupant	Geofency False Negative as a Percentage of Occupied Time	Geofency False Positive as a Percentage of Unoccupied Time	Wi-Fi False Negative as a Percentage of Occupied Time	Wi-Fi False Positive as a Percentage of Unoccupied Time
Person B	24.6%	0.6%	46.5%	0.7%
Person D	0.2%	1.3%	81.7%	1.4%
Person F	6.8%	0.3%	56.8%	0.8%
Person J	28.9%	7.9%	28.4%	5.3%
Person K	4.2%	1.9%	14.8%	2.7%
Person P	6.1%	0.2%	25.9%	0.7%
Average	**11.8%**	**2.0%**	**42.3%**	**1.9%**

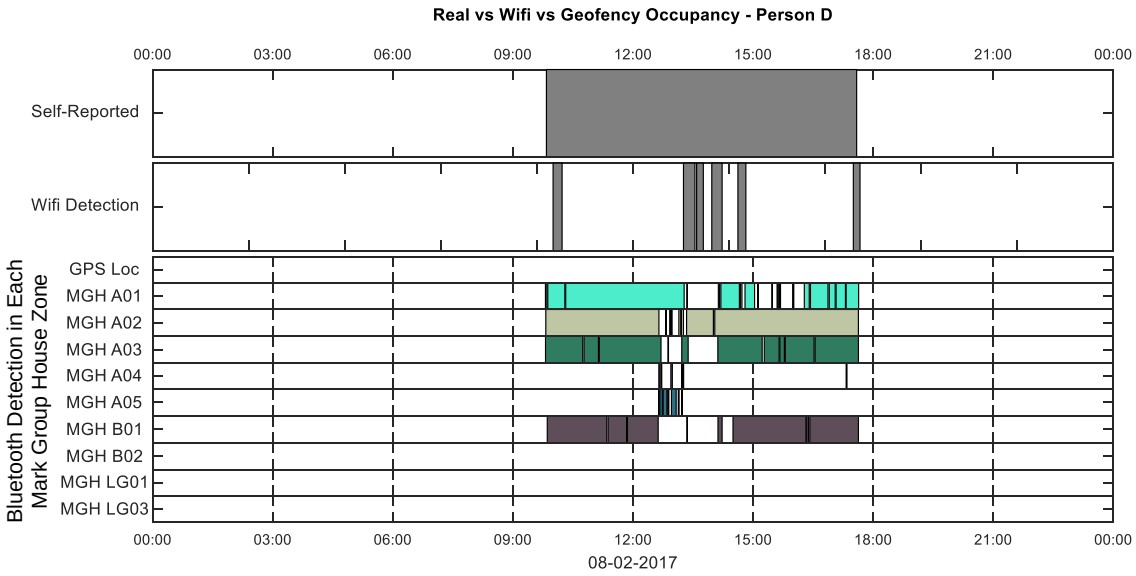


Figure 5-21 - Geofency Results vs Wi-Fi Detection and Self-Reported Presence - Phase 2 Kontakt iBeacons

5.4.5 iBeacon Software Development – Android

An app equivalent to the iOS Geofency was not found for Android users – with most options available not able to run in the background over a long period of time. It was decided that the best solution would be to use open-source libraries available for Android development to produce an app that met the testing requirements. The requirements for the Android iBeacon Localisation software were as follows:

- Scans using Bluetooth-LE for nearby iBeacons from a predefined set of beacon hardware.
- Scanning is run on a continuous background service, even when the app is not active.
- Creates a log of when the phone is within range of each of the localised beacons.
- Allows visual review of the logged data for manual verification by the user.
- Where multiple beacons are detected at once, the closest beacon is identified as the current location.
- Noise reduction on data where necessary.

The Bluetooth-LE scanning app was developed in Android Studio [199] using the Java programming language. A plain-language version of the algorithms used in the app is included in Appendix 10.5, with the full Android Studio project included in the physical copy of this work in Appendix 10.7.

The AltBeacon Android Library [200] is a widely used and well documented library available for Android development. By default it is compatible with the AltBeacon standard, but can be configured to detect any generic iBeacon signal [201]. The app was developed with a similar structure to the open-source project described in the article above, with a background service running periodic Bluetooth scans for local iBeacons.

During each periodic scan for beacons, a list of visible beacons and their received signal strength is generated. From this list, the closest beacon is calculated every thirty seconds. When the closest beacon changes or is not found, this is logged to a file stored on the device's internal storage, which can be reviewed via the app or sent out by email.

5.4.5.1 Testing and Troubleshooting

The app was initially tested in a simplified 3-zone setting over a single storey, with two regular occupants. The location of one occupant was recorded manually to an accuracy of one minute. Figure 5-22 shows the manual 'ground truth' data, while Figure 5-23 shows the data collected by the Android app. It can be seen that the location estimation is accurate for the majority of the testing period.

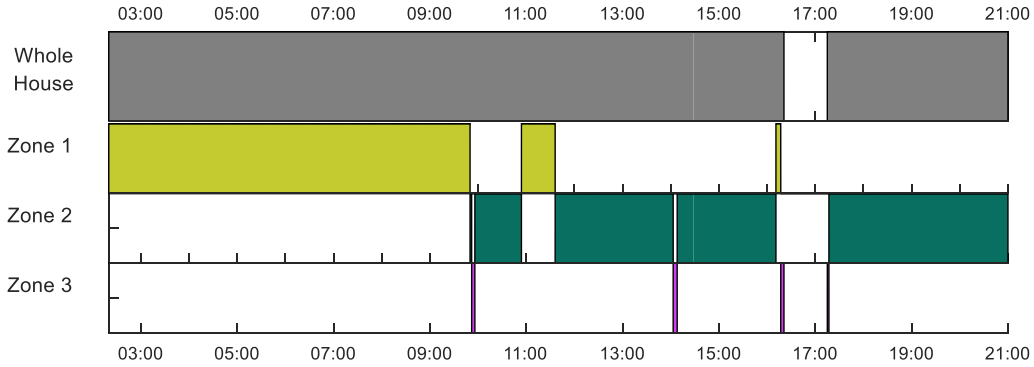


Figure 5-22 - Ground Truth Occupancy Data for Test Period

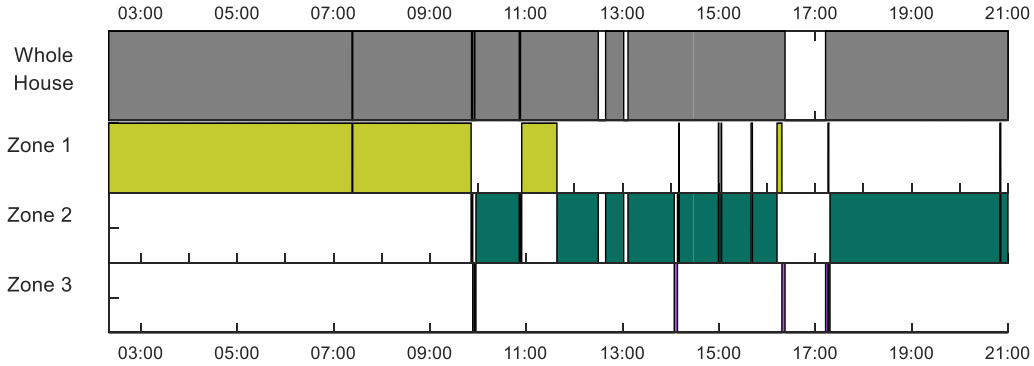


Figure 5-23 - App-Collected Data for Test Period

Figure 5-24 shows a visualisation of the app performance against the ground truth data for each zone, and a summary of the accuracy of binary presence (whether the occupant is present in any zone over the whole house) over the same period. As shown, the overall accuracy on this simplified case is high, with an overall accuracy of 98% for binary occupancy, and each individual zone between 97% and 99%. The rate of false negatives is higher than the rate of false positives. This is to be expected: false negatives can occur when the service is interrupted or the signal is not received due to interference etc. False positives, on the other hand, should not be possible except in the case where multiple zones are seen at once and the closest one is misidentified. In theory, a presence signal from the beacon cannot be received unless the occupant is actually present, so the whole-house binary presence graph should

show a zero false positive rate. The brief false positive shown on the binary occupancy graph can be explained due to a slight rounding inaccuracy in the ground truth reporting.

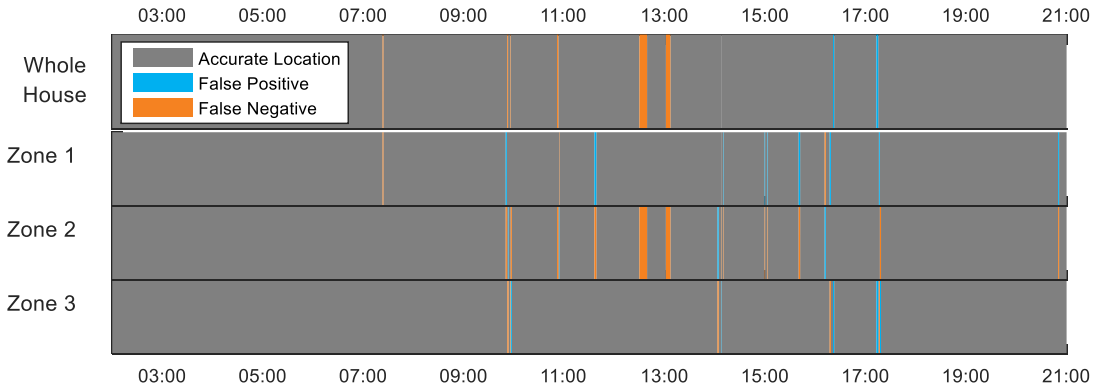


Figure 5-24 False Positive and False Negatives for each zone over the test period

After showing promising results in the simplified test, the app was tested on the same mobile device in the Mark Group House, with one beacon per zone over three storeys of the building, and up to 15 occupants present in the building during tests. The results were found to be much more variable over several tests, with accuracy of location ranging from 0% to >60% for the same location on different days. An example test is shown in Figure 5-25: while the user was static in Room A02, the app logged their location in 4 different zones over the tested period.

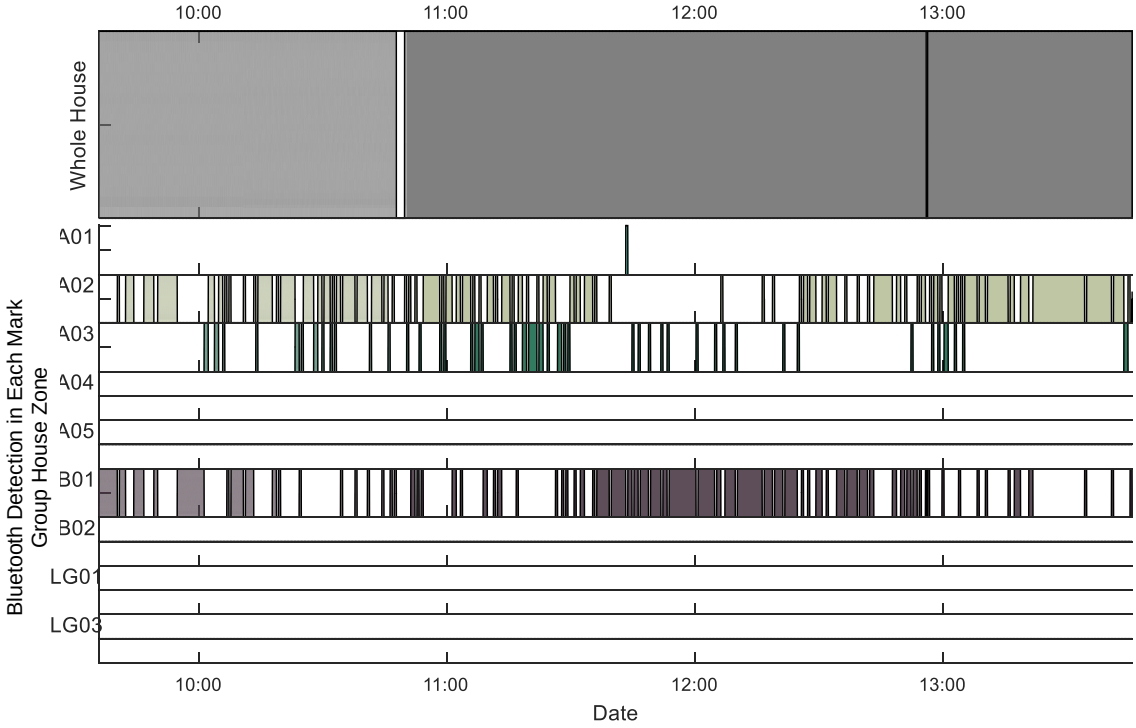


Figure 5-25 App-collected data from the Mark Group House while User was in Room A02

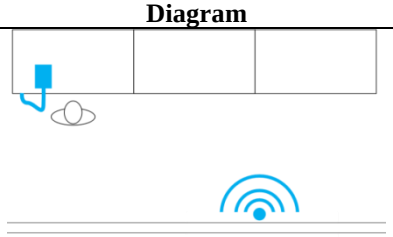
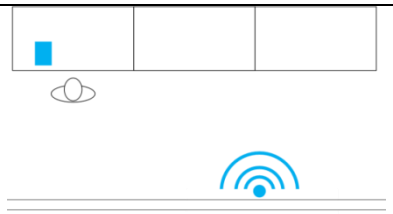
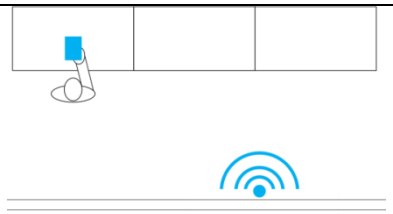
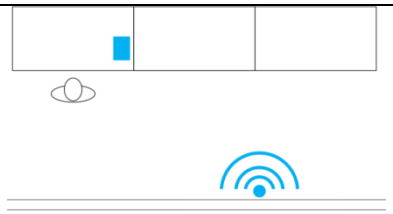
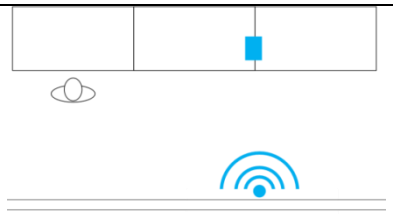
After observing the conditions of the building over several tests (summarised in Table 5-8), it was found that the number of occupants present around the mobile device and the location of the device relative to its owner appeared to have an effect on the beacon signal strength received. This was tested more formally by recording the location of a device during different conditions within the same zone. The configurations tested are described in Table 5-9.

Table 5-8 - Summary of Several Software Tests During Occupied Hours

	Percentage of time identifying correct zone	**Occupants present local to device**
Test 1	41.40%	Single occupant
Test 2	84.10%	Single occupant
Test 3	3.20%	Additional occupant sitting between beacon and phone
Test 4	7.50%	Additional occupant sitting between beacon and phone
Test 5	0%	Single occupant, device held by occupant

In this test, a third-party application was used that creates a log of each individual beacon detection and its received signal strength (RSSI) [202]. This application can only function while running in the foreground of a device, and so was not suitable for use over longer periods of time.

Table 5-9 - Occupant-device configurations tested for their effect on signal strength

Test Period	Condition	Diagram
A - 15:50-15:55	Mobile device is plugged in to PC for charging. Device is placed with a human body in the direct line of sight from beacon to device.	
B - 15:55-16:00	Mobile device is kept in the same location as previously, but unplugged from the PC.	
C - 16:00-16:05	Mobile device is held in hand, but otherwise in the same location.	
D - 16:05-16:10	Device is moved away from the user, approximately 1m closer to the beacon. The user's body is no longer between the beacon and device.	
E - 16:10-16:15	Device is moved to approximately 1m away from the beacon, with no physical obstructions between the two.	

The received signal strength graph in Figure 5-26 shows that the local conditions of the receiving device have a significant effect on the signal strength received from each beacon, even so far as to change the perceived order of closeness:

- Objects touching the receiving device can affect its received signal strength – in sections A and C the range of signal strengths received is increased while the device is in contact with a charging wire or the human body. Closer beacons are detected with increased signal strength, suggesting that another

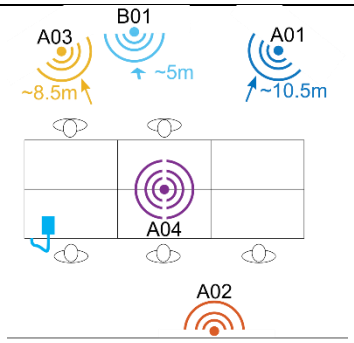
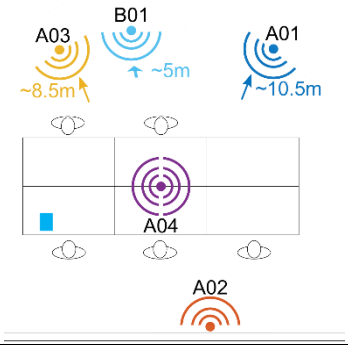
body in physical contact with the device may be acting as a boosted receiver for the Bluetooth signal.

- The beacon in the room above the tested zone (purple markers, 7643-22828) shows step changes in the strength of signal between sections B, D and E, despite only a small change in actual distance to this beacon. The lowest received strength occurs in section E, when the device is placed closer to the beacon than in section B or D. This suggests that the materials of the building structure or furniture can have a significant effect on the signal strength, with some obstruction between the device and beacon present in location E.
- The beacon of the tested zone (red markers, 23429-1275) has the lowest average signal strength when a person sits between the device and beacon (section B).
- When the device is placed furthest from the person, the signal strength received for the closest beacons becomes more consistent – this may suggest that the small movements made by an occupant can affect the signal strength in between each detection.

Figure 5-26 - Signal Strength Received by Device in Test under Different Conditions

It was then investigated how effectively the received signal strength was translated to the closest reported zone by the location app. The developed app was run concurrently with the third-party scanner used in the test above, and the results compared. A second beacon was added to the test zone (see Table 5-10), to measure whether a more central location could reduce interference from bodies between the device and beacon.

Table 5-10 - Occupant-device configurations tested with two local beacons

Test Period	Condition	Diagram
F - Figure 5-27	Mobile device is plugged in to PC for charging. Five occupants present in zone, seated as shown. Two beacons placed in zone.	
G - Figure 5-28	Device is placed clear of lower left occupant. Five occupants present in zone, seated as shown. Two beacons placed in zone.	

The distance of each beacon can be calculated using the RSSI and the expected power level at 1m (Tx). The formula used by the Altbeacon library [203] is:

$$d = A\left(\frac{RSSI}{Tx}\right)^{B} + C$$

Where *d* is the estimated distance from the beacon in metres, *RSSI* is the received signal strength in dbm, *Tx* is the expected signal strength in dbm at a reference distance of 1m (this value being beacon-specific and broadcast as part of the beacon signal by the beacon hardware), and *A*, *B* and *C* are constants determined empirically for different models of mobile phone [200]. Figure 5-27 shows the signal strength calculated manually from the RSSI data, using the AltBeacon library data for a Motorola device, as well as the reported closest beacon from the developed app.

In theory, the A04 beacon is the closest to the occupants sitting at the desks, and should not be affected by signal attenuation from human bodies. It should be noted during the test, there was significant movement of occupants around the building. It was also noticed that, while the phone was not placed behind any full bodies, the occupant at the lower left desk blocked the direct line of sight to the beacon with their arms while typing, which occurred intermittently through the experiment. This may have had an effect on the received signal strength.

The results in Figure 5-27 show that the signal strength received is still highly variable – such that during the half-hour test period, the four of the five visible beacons had similar amounts of time identified as the closest. Beacon A02 was logged for 27.6% of the tested period. A04 was logged for 32.6% of the tested period.

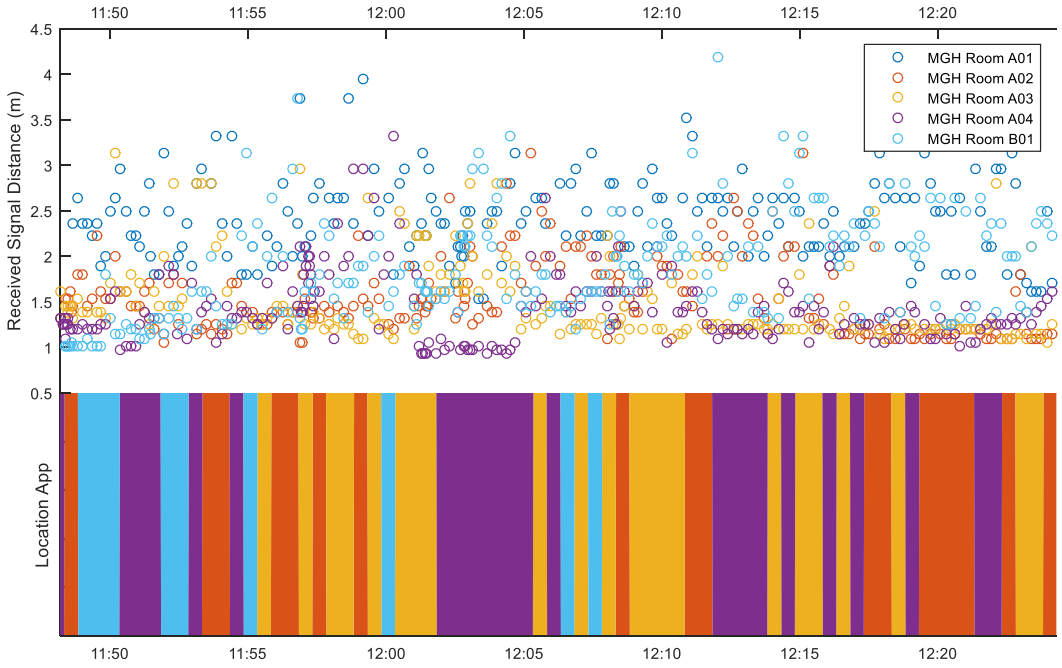


Figure 5-27 - Third-Party App and Location App Detections during test F

With the device placed out of range of the occupant at the first desk (Figure 5-28), the results show that the two closest beacons were always correctly identified when the phone was placed away from the occupant's arms, suggesting that even the arms of an occupant can attenuate the signal enough to affect the closest detected beacon.

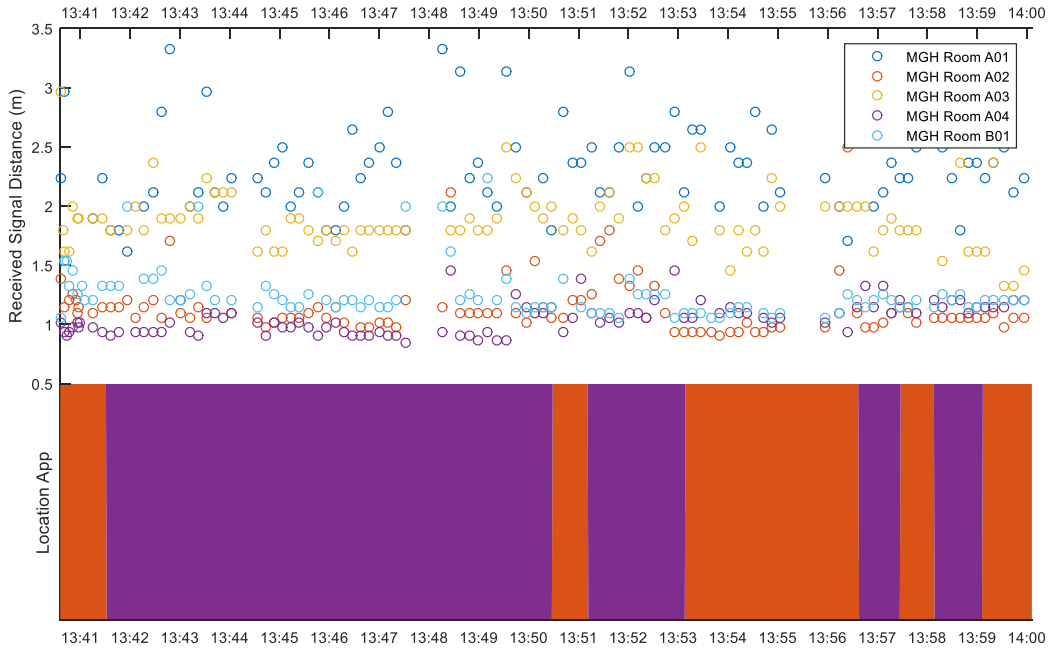


Figure 5-28 - Third-Party App and Location App Detections for test G

The app results were also tested as a measurement of binary occupant presence, in a similar manner to the testing for the iOS app in Section 5.4.4. One week of manually recorded location data was used as a reference for detection accuracy. When implemented across six Android devices for the test week, it was found that the Android app suffered a dramatically reduced performance on some occupants' devices. While the app remained active on all but one device, it was found that devices running the 'stock' Android operating system (persons A and O) kept receiving Bluetooth signals while the device was not active, and devices running manufacturer-modified versions of Android (persons E, L and Q) did not receive any Bluetooth signals while the app was running in the background. This suggests that Android device manufacturers may apply more aggressive system checks to shut down background processes using Bluetooth connections – a problem that may not be possible to overcome on all device types. Table 5-11 demonstrates how the overall accuracy can be a misleading measure of success: the average accuracy for the app is over 90% despite some occupants receiving no Bluetooth signals over the test week, due to the relatively short periods of occupancy over a full 24h period, estimating occupancy as a constant zero can yield a high accuracy, although it is clearly not a successful detection method. A better measure of success is shown in Table 5-12, where the false negative rate as a percentage of self-reported occupied time is

presented. This measures how often the app 'misses' an occupant while they are actually present. For occupants A and O, it can be seen that the app outperforms the Wi-Fi detection in this measurement. For the other occupants, it can be seen that their presence is completely missed. For occupant S, the results are poor, although it should be noted that this was due to issues with the app crashing and losing data for the first 6 days of the test week. While the app was functional, the successful detection rate appeared to be similar to occupants A and O, as shown in Figure 5-29.

Table 5-11 - Summary of Android iBeacon App vs Wi-Fi Agreement per Occupant

Occupant	Wi-Fi-Android App Agreement	Wi-Fi-Android App Agreement while assumed occupied	Android App Accuracy with Self-reported Occupancy	Wi-Fi Accuracy with Self-reported Occupancy
Person A	97.1%	86.2%	99.4%	96.9%
Person E	92.5%	0.0%	90.6%	93.8%
Person L	93.0%	0.0%	86.6%	91.1%
Person O	92.8%	30.6%	99.7%	92.9%
Person Q	86.6%	0.0%	85.2%	96.7%
Person S	83.8%	19.3%	84.1%	93.5%
Average	**90.9%**	**22.7%**	**90.9%**	**94.1%**

Table 5-12 - Summary of False Negative/Positive Rates for Android iBeacon App and Wi-Fi Detection

Occupant	App False Negative as a Percentage of Occupied Time	App False Positive as a Percentage of Unoccupied Time	Wi-Fi False Negative as a Percentage of Occupied Time	Wi-Fi False Positive as a Percentage of Unoccupied Time
Person A	1.4%	0.4%	11.8%	0.9%
Person E	100.0%	0.0%	43.2%	2.4%
Person L	100.0%	0.0%	56.8%	1.4%
Person O	1.5%	0.2%	66.5%	0.8%
Person Q	100.0%	0.0%	15.8%	1.2%
Person S	78.8%	0.1%	17.5%	3.8%
Average	**63.6%**	**0.1%**	**35.3%**	**1.7%**

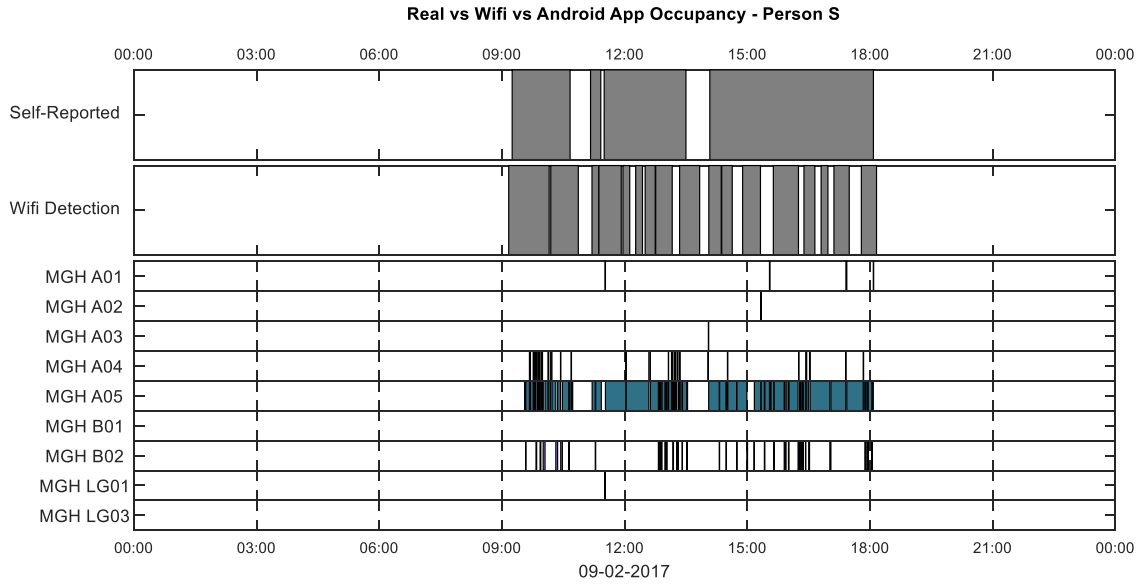


Figure 5-29 – Android App Results vs Wi-Fi Detection and Self-Reported Presence – Person S, final day

5.5 Proposed Framework for Inferring Occupant Information

The work above has set out the methods applied to collect a wide range of occupant-centred data. However, for this data to be useable, there needs to be a further step to parse raw data into streams of information that are more directly relevant to the intended application. As discussed in previous sections, this work focuses on a system for the short-term prediction of localised occupancy rates, identified through the review of existing commercial BEMS and applied research in Chapters 2 and 3.

To make the most effective control decisions, the system must be able to detect all types of occupants within a space, including non-regular occupants and those not carrying personal devices. This suggests that, while the device detection methods tested in this chapter show promise as standalone systems, a combination of data sources is likely necessary for a more comprehensive solution. In the study of existing work, it was found that there is little standardisation of occupancy data collection within buildings, highlighting the need for a more systematic way to assess what is needed and how to structure data collection/processing systems.

The proposed predictive system is effectively split into two tasks: being able to infer local occupancy rates from raw sensor data, and then using inferred rates to predict near-future changes. There are a number of benefits to deliberately separating these tasks into two independently run, but interacting systems:

- Recent years have shown rapid advances in the technologies available for indoor occupant data collection. Separation of detection and prediction allows for updates and improvements to the technologies used for detection without requiring a new predictive model to be built.
- Frequency of updating required is likely to be different for the two tasks: future prediction models need to respond quickly to changing behavioural patterns and so should be regularly updated, while this is not necessarily the case for the relationship between sensor data and occupancy rate.
- The ability to draw and use current occupancy state separately to the predictive outcome: while the predictive future occupancy rate is useful for slow-response systems, the current occupancy state alone could be drawn to use in faster-response systems or wider building monitoring uses.
- Potential for operating at different time resolutions: some fast-response systems such as lighting require immediate feedback when occupancy is detected, while it would be unnecessary to re-calculate a one-hour future prediction on the same timeframe. Splitting the two tasks allows the detection and prediction of occupants to be operated at time resolutions appropriate to each task.

Given these benefits, in this work the system to parse occupant information is split into two interacting, but distinct modules as shown in Figure 5-30. This figure illustrates how the occupancy modelling modules could function as part of a greater building monitoring and control system, with a centralised information processing system able to output information to the building controls and other analytics as needed.

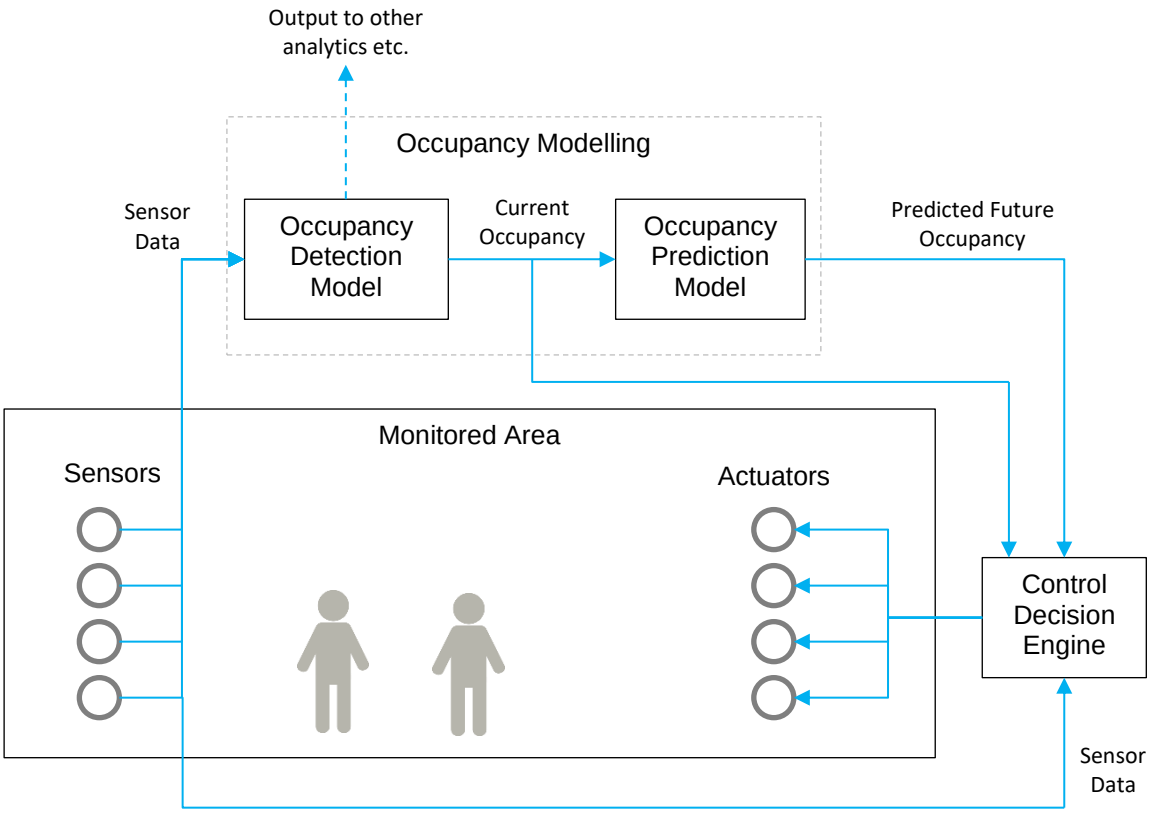


Figure 5-30 - Diagram of Proposed Occupancy Modelling Structure

5.6 Conclusions

In this chapter, systems aimed at minimal-cost, opportunistic collection of occupant data using wireless signals and users' personal devices were set up and tested. Two methods were pursued: the collection of existing Wi-Fi signals from personal devices, and the use of personal devices to listen for locally placed Bluetooth LE beacons. Both systems were tested against manual recording of local occupancy rates over a limited time period.

The system set up to detect probe signals from Wi-Fi devices was able to provide valuable information on building occupancy trends over time. The detection of devices from Wi-Fi traffic can collect more generic information than counting the number of devices connected to a specific network, as this system also counts visitors who do not actively connect to any local Wi-Fi networks. Devices were found to not be detected continuously, however 80-90% of continuous detections from the same device were seen to occur within a 5-10 minute window. This number, however, was highly variable by device, with some devices consistently showing a number of hours between probe signals.

The signal strength of the received Wi-Fi signals was found to have a relation to locality, with a significant drop in signal strength seen caused by both distance and blocking by building fabric. This allows the signal strength to be used to filter out false positive results from people passing nearby the monitored area, but this is dependent on the location of the sensor and makeup of the building. In spaces with a large population, the trend in the number of devices seen gives potentially more information than the actual number of devices, which was not equal to the estimated local population.

Bluetooth beacon based systems can provide localised data with easier installation and lower hardware costs than listening for Wi-Fi probe signals, but such systems require data collection software to be installed and run on participants' own devices. The use of Bluetooth-LE specifically allows for more frequent probes, meaning that finer-grade data can be collected with less of an impact on the battery life of components. However, signal strength of Bluetooth was found to be highly variable due to physical properties of Bluetooth frequency waves: the signal was easily blocked by people, furniture and building fabric. This was seen to significantly skew results when comparing the closeness of multiple beacons. It was also found that some personal devices had more aggressive energy saving software, turning off active processes for Bluetooth scanning while the device was asleep. This meant that some of the participants produced no useable data during the test period. For those that did collect data, binary presence accuracy of the BT system was high, while location accuracy varied depending on the local environment.

Where issues were found with the physical signal attenuation using the Bluetooth LE beacons, there is potential for improvement with recent advances in technology. The Bluetooth 5.0 protocol was released in late 2016, with devices supporting the protocol releasing through 2017. This Bluetooth update has been reported to increase the range of signal by up to four times the previous Bluetooth 4.2 specification [204]: this could reduce some of the signal strength issues seen in this chapter and may allow a Bluetooth-based solution to provide more reliable indoor location across a wider range of applications.

In general, it was found that both tested systems for a low-cost solution to detect personal mobile devices saw mixed success, and would not be appropriate for use

alone in an occupancy detection system where accuracy and reliability are important. As part of a greater system, these two measures may provide valuable information. As stated in previous chapters, for the purpose of control decision making, it is also important that all building occupants are detected: a diverse set of data sources are needed to ensure that occupants who may not be carrying mobile phones, may not interact with access control systems etc. are still counted.

Following from these conclusions, a framework for parsing useable occupancy information from raw data sources was proposed. Within this framework, a modular approach was chosen to allow for greater flexibility to changing sensor technologies, the possibility of delivering occupancy information to different systems at time resolutions appropriate to each application and the ability to update parts of the occupancy model independently as required. The following chapters will focus on developing and testing models for these two tasks. Chapter 6 covers the combination of data from multiple sensors with the aim of improving the overall accuracy and reliability of occupant detection beyond the capability of the systems developed in this chapter when applied alone. Chapter 7 covers the development of a model to predict near-future occupancy rates based on the inferred occupancy level from the previous module.

6 OCCUPANCY DETECTION MODEL DEVELOPMENT

6.1 Introduction & Aims

As discussed in the review of existing work on occupancy detection, it was found that short-term prediction of future occupancy can allow for the greatest energy saving in pre-emptive control of systems. As discussed in the previous chapter, the task of converting sensor data to a prediction of the number of occupants in a space has been split into two modules, to allow for changes in available technology. The first module is a model to convert sensor data into an estimation of the number of people currently in a space in real time. The second module is a model to take the number of people over a recent time period and predict the number of people in the future.

This chapter covers the first stage of this model – 'occupancy detection', designed to convert a range of sensor data from environmental and personal devices detection sources into a localised measure of the number of people in a space. The development of this occupancy detection model included as wide a range of occupancy data as possible in order to compare the value of data from each source, test how much information can be gained in the combination of data sources and highlight any redundancies that could allow for more efficient data collection in future applications.

A full description of the small office building tested can be found in section 5.2. The data used in this chapter includes environmental sensors for motion, CO_2, temperature and humidity in each zone, external door/window interactions and the Wi-Fi and Bluetooth data collected through the systems described in Chapter 5. Ground truth values used for training were collected by manual surveying of occupant location over a test period. The range of sensors implemented was designed to achieve the greatest range of occupancy data while minimising disruption to the building and maintaining privacy for any occupants who have not specifically opted in to being personally tracked by the system.

6.2 Selection of Appropriate Machine Learning Methods

Given the wide range of data sources available, many of which are interdependent, it is unlikely that the occupancy rate can be well represented by a simple set of heuristics. The application of machine learning techniques offers an automated way to learn the complex interactions between sensor data and occupancy rate.

One of the initial choices to be made was whether the model would be based on classification or regression. In the case of a model estimating the number of people in a space, either model type could be feasible. Figure 6-1 illustrates the difference in how the model output is represented for classification a) and regression b). For classification problems, the output of the model is constrained to a set of pre-defined categories. Given sensor data inputs, a classification occupancy model would output the most likely category given the input data: in the illustrated case, 3 occupants in the zone. For a regression model, the estimated number of people could be any number, including non-integer values, based on a modelled relationship between the value of the sensor input and the value of the number of people during training. All of the methods described below are 'supervised', using a set of labelled training data or 'ground truth', where a number of examples of the output variable with their corresponding input variables are provided. Non-trained 'unsupervised' techniques are typically used for identifying recurring patterns or clustering data into similar groups, which then must be manually interpreted into actionable situations by a human reader, but would typically not be able to interpret actual numbers of occupants.

Figure 6-1 - Illustration of a) Classification and b) Regression Based Model Outputs

For spaces where the number of occupants is known to always be within a certain, limited range, a classification model will likely be more effective. For example, the work of Yang et al [108] concluded that classification, in this case especially using Decision Tree methods, was more effective for both single and multi-occupancy spaces. In the above study, the maximum number of occupants in a space was predefined and was not exceeded during the data collection period.

However, in this study, the building tested has several spaces that can be occupied by an unpredictable number of people, with large groups of visitors occasionally occupying a demonstration space or mixed groups of known and unknown occupants

filling a meeting space beyond its usual capacity for a short meeting or presentation. More importantly, it is unlikely that the manually labelled training data over one week will cover all possible numbers of occupants for each space. For a classification model, this behaviour represents a challenge: any categories without training data cannot be correctly modelled, and the number of people in a space must be assigned a definite limit. For these reasons, it was decided that, for this study, a regression model trained on the limited 'ground truth' data available, with its ability to interpolate from known numbers of occupants, would be more effective at identifying reasonable occupancy rates, in particular in identifying occupancy levels that fall beyond the scope of the initial training dataset.

There are several machine learning techniques that can be used for supervised regression problems. The methods considered are summarised briefly below.

6.2.1 Statistical Regression Methods

Linear Regression

The linear regression method is generally applied in cases with a less complex relationship between the input variables and the output. Linear regression assumes this relationship takes a linear form [205], for example:

$$f(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 \quad (1)$$

Where x is the vector of input variables (in the case of this study, the sensor data) and f(x) is the prediction of the output variable. The objective is then to find the values of the θ parameters that most closely satisfy the training data. This can be achieved using one of a range of optimisation techniques, which will be discussed in a later section. Linear Regression is not considered appropriate for this application, as the relationships between sensor data and the number of local occupants will likely not be well described with linear relationships.

Polynomial regression

Polynomial regression uses the same process as linear regression, but the assumed relationship takes a polynomial form, rather than linear [206]. In the case of a complex relationship between inputs and outputs, this method is more likely to be useful than linear regression, but needs the order of polynomial to be predefined. Given that in this case, the nature of the relationship between the sensor data and

number of occupants is largely unknown, but is likely to be highly nonlinear, polynomial regression may not be the most appropriate method.

One of the major considerations when setting up most types of supervised learning models is the issue of over or underfitting to the training data. In the case of polynomial regression this typically takes the form of choosing the wrong order of polynomial, and is illustrated in Figure 6-2. Underfitting issues arise when the model is not capable of properly representing the complexities of the training data, while overfitting is seen when the model is too complex for the amount of data supplied and finds patterns or correlations in small errors in the training data. Overfitting is characterised by low error on the training data, but high error on new examples.

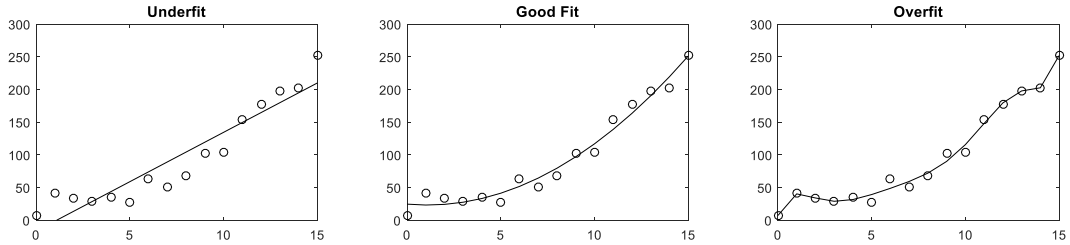


Figure 6-2 - Illustration of Polynomial Underfitting and Overfitting Issues

6.2.2 Instance-based Learning

K-Nearest Neighbours (kNN)

One of the simpler machine learning methods based on labelled training data, k-Nearest Neighbours is a non-parametric method based on the most similar available examples from the training set [207]. The kNN method does not explicitly train a generalised model, meaning that it can be quick to set up. Based on its input variables, the output value of a new example is taken as the mean of the *k* nearest examples from the training data, where *k* is a number chosen to optimise the fit of the model.

As this method works on proximity to previously seen examples, it does not give good performance outside of the scope of the original training data. In the case of estimating the number of occupants per room with a limited amount of training data, this causes significant problems, as it cannot be expected that all rooms will encounter a full range of possible occupant numbers within the training data period.

Decision Tree/Random Forest

Decision trees are generally used for classification problems, but can also be applied to regression. This method was found to be the most effective for a classification-based occupancy model in a comparison of different machine learning techniques [108]. As a non-parametric method, decision trees are able to represent highly nonlinear relationships without manual specification of the expected complexity of the model [208]. At each branch, the decision tree divides the training examples using some criteria from their input variables (for example, dividing cases where PIR count > 0 from those where PIR count = 0), with the aim of maximising the variation in the output variable at each division. This produces a model structure as shown in Figure 6-3, where a new example is sorted by travelling down the tree structure according to the values of its input variables, with its class or value estimated by the most common class or mean value of training examples at the end node.

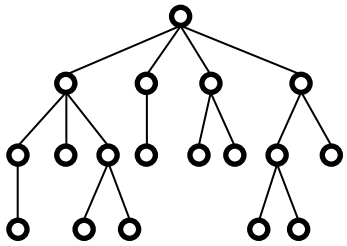


Figure 6-3 - Illustration of Decision Tree Model Structure

As the tree structure divides data based on the value of the input variables, it is typically more suited to discrete or categorical inputs than continuous variables, with which it can be more difficult to draw definite boundaries. This would potentially be a problem with the sensor data for this study, which has both continuous (CO_2, temperature) and discrete (PIR, Wi-Fi presence) inputs. As with kNN, this method is also not able to predict values beyond the scope of the training data, as its outputs are based on a mean of examples that have been seen before. This method was therefore considered unsuitable for application in this study.

The random forest method uses multiple decision tree models trained on different subsets of the training data. Each tree's output counts as a 'vote', with the mode (for classification) or mean (for regression) output of all trees taken as the estimated value of a new example. This method can reduce the issues with overfitting found with a single decision tree, but still suffers from the characteristics described above that

make the method unsuitable for this application. However, ensemble training methods similar to the random forest can be applied to other machine learning types.

6.2.3 Support Vector Machine Regression (SVM)

As a classification tool, SVMs seek to draw a linear boundary between classes that maximises the distance between the training examples and the decision boundary, producing a more robust model [209]. With the use of kernels this method can represent highly nonlinear systems without the need to specify polynomial parameters. This is achieved by mapping the original input variables to higher-dimensional features using kernel functions, such that a linear solution can be found for these higher-dimensional features. A simplified case is illustrated in Figure 6-4, where two classes *x* and *o* are represented by input variables i_1 and i_2. The classes cannot be separated using linear or low-order polynomial methods, when the inputs are mapped to a feature f_1 using a Gaussian kernel function centred at point A, the boundary can be represented by a linear plane in f_1.

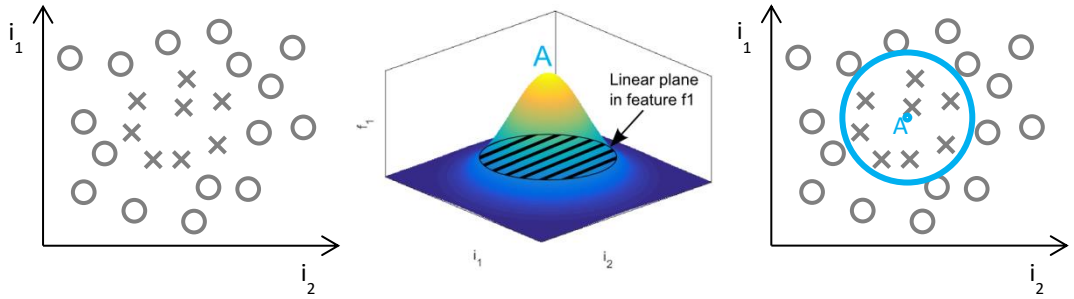


Figure 6-4 - Illustration of Kernel-based Classification using SVM

SVM kernel methods can also be applied to regression problems using a similar process to map inputs onto features that can be represented using linear regression. As this method is appropriate for highly nonlinear systems and can give robust solutions, it was considered a potential candidate for application in this study.

6.2.4 Bayesian Methods

Bayesian methods apply Bayes' theorem (equation (2)) with the aim of outputting a probability distribution across possible outputs *y* given a set of input values *x*, rather than a single value output as in other machine learning methods. The estimated value of the output variable can then be taken as the highest probability value of *y*, but some consideration can be given to the certainty of this output using the output probability distribution itself.

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} \quad (2)$$

The following methods apply Bayes' theorem directly in their 'core' versions, but it should be noted that several non-probabilistic machine learning models can also be modified to include Bayesian techniques.

Naïve Bayes

Naïve Bayes models are typically used for classification problems, but can also be applied to regression [210]. These models are generally favoured in cases that require a simple structure and quick implementation. A model for the probability of observed input values *x* given the known output *y* is constructed using a labelled set of 'training' data. The major assumption applied to this model type is that each of the input features in *x* can be considered to be independent given *y*. In the case of occupancy detection using environmental sensors, this may not be a valid assumption, as some variables depend on external influences such as the external weather conditions and some variables are directly related, such as the temperature and relative humidity. Some studies also show that Naïve Bayes is less effective in regression problems than for classification [210]. It was decided that a Naïve Bayes model would not be the best solution for this application.

Gaussian Process

The Gaussian Process regression method seeks to find the probability of a function *f(x)* that describes the output *y* when given inputs *x*. The 'prior' in this case is to assume a probability for a theoretically infinite range of possible functions, and refine the probability of this range based on the observed training data, as illustrated in Figure 6-5. It can be seen that the range of possible functions drawn from the posterior is limited to those that fit the observed data, meaning the model becomes more certain close to observed data points.

Similarly to SVMs, Gaussian Process models can use a range of kernel functions that can affect the outcome of the model. In this case, the kernel or covariance function of the Gaussian Process prior is used to encode some expectations about function *f(x)*, although not its exact parameters or structure. For example, certain hyperparameters of the covariance function can affect the 'smoothness' of functions drawn from the

prior, which could be used to tune over or underfitting issues. The values of the covariance function hyperparameters can be determined from the training data using methods described in the work of Rasmussen and Williams [211].

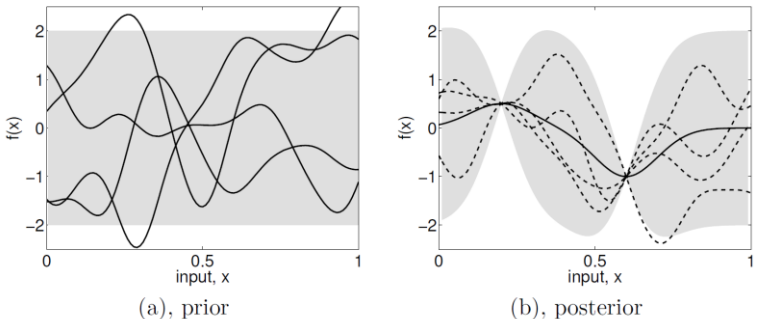


Figure 6-5 - Visualisation of Gaussian Process functions drawn from the prior distribution before data (a) and posterior distribution after inclusion of training data (b) [211]

One of the major advantages to the Gaussian Process method is the ability to extract a level of certainty in the model outputs. In the case of reducing uncertainty around building occupancy, a level of confidence in the number of people could be a useful parameter to feed into controls systems. Gaussian Processes were therefore considered for application in this study.

6.2.5 Artificial Neural Networks (ANN)

Artificial Neural Networks, also called multi-layer perceptrons, are designed to represent the relationship between a set of inputs *x* and the output *y* without explicit knowledge of what form this relationship will take. This is achieved by mimicking the way that neurons in the brain function: representing the model as a network of nodes or 'neurons', linked by pathways which are assigned weights *w* [212]. Each of the individual inputs is represented as a node on one side of the network and the output(s) on the other side, as shown in Figure 6-6. A number of 'hidden' nodes are arranged in layers between the inputs and outputs. Each node has an input function, activation function and output such that a set of input values *x* are passed through each layer, being altered by the weights of each node connection, until an estimated value of the output *y* is calculated at the output layer.

The values of the weights *w* that give the best estimated output must be found by optimisation using labelled training data, for which the inputs and outputs are known. Various optimisation algorithms can be used for this training period, with the aim to find weights to minimise the 'cost function', which is a measure of the total error in the output over the training dataset as a function of the weight values.

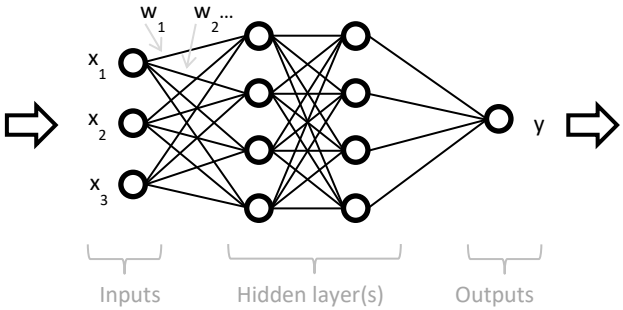


Figure 6-6 - Simplified ANN Structure Diagram

ANNs have been found to be a highly effective tool for regression in direct comparisons of different machine learning methods on time-series data [213]. The ability to represent highly nonlinear relationships without the need to predetermine a polynomial structure or evaluate high-order polynomial terms is a benefit for application in this study. ANNs were therefore considered as a potential solution for application.

6.2.6 Discussion

From the array of possible machine learning algorithms available, several viable model types were identified: support vector regression, Gaussian processes and neural networks. As each of these model types has no outright failings when considered for this application, to some extent the model type used is less important than the way the model is refined once initial tests have been made on the dataset. After consideration, an Artificial Neural Network model was selected, given that:

- ANNs are able to represent nonlinear systems without prior knowledge of the complexity of the problem – in the case of a large number of sensor inputs with some degree of dependency on non-occupancy factors, it is expected that the relationship will be highly nonlinear and potentially noisy.
- Has been shown to be effective in studies directly comparing methods: while this is no guarantee that ANN would be the most effective method in this particular application, the general characteristics shown by ANN models suggest they should be appropriate.
- The availability of support resources for this model type was found to be more robust in the chosen programming language than other model types.

It should be noted that, as a supervised learning method, ANNs rely on manually labelled training data to build the relationship between the input values and the output

estimation. As was discussed in Chapter 5, two periods of manually recorded building occupancy data were collected for this purpose. During the course of this collection, it was found that some occupants reported to a greater degree of precision than others, and some occupants reported rounding time values as they were unsure of the exact time of transitions. For the purpose of this study, it must be assumed that the training data used represents the true occupancy of the space, and the model output will be compared to this value as if it is 100% accurate. Where this assumption may have affected the results of the model, this will be discussed in the following sections, although it is expected that the overall effect on this work is minimal.

6.3 Model Structure

A range of software for the implementation of ANN is available. In this study, it was chosen to use the Matlab Neural Network toolbox, due to:

- Availability of functions within this software.
- Researcher familiarity with the programming language.
- Existing scripts for the analysis of occupancy data in Matlab for the other case studies in this project.

Figure 6-7 shows the initially tested structure for the detection model. In this diagram, each circle represents a neuron, with the lines between neurons representing the weights that are optimised during model training. During the first stages of testing, one such model was trained per zone of the building, as the number of people in a zone with a given set of sensors should be independent of the occupancy of other zones. Training individual models also allows for a quicker way to test the effectiveness of different structures, data inputs etc. on each room type. In real application, the separate models could be combined into a single model structure, with the inputs of each zone separated from each other, or the models could be run one at a time over the 5-minute update period of a real-time system.

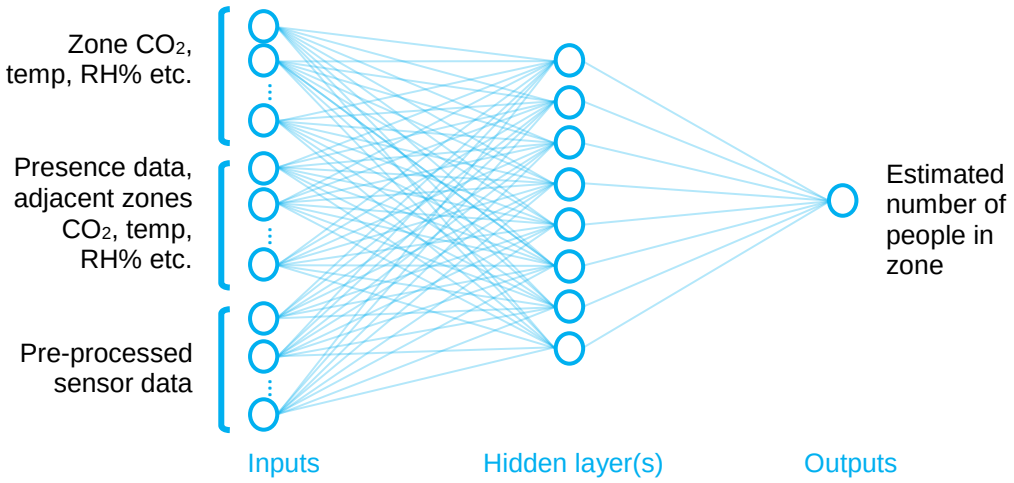


Figure 6-7 – Initial Tested ANN structure for Detection Model

When training the ANN model, a large number of variables can affect the likelihood to over/underfit to data, including:

- Number of neurons in hidden layer(s)
- Number of hidden layers
- Number of data inputs
- Format of data inputs
- Format of data outputs

To attempt to try every permutation of each of these variables in order to optimise the ANN performance is not computationally viable. Some logical rules can be applied to eliminate combinations that are extremely unlikely to yield good results. For example, introducing a high number of data inputs to a network with a low number of neurons is unlikely to be able to represent any complex relationships between inputs. Conversely, training a network with few inputs, with a high number of hidden layers, each with many neurons, will overfit the network to the training data.

Given the limited amount of training data, a highly complex ANN structure is not feasible without introducing overfitting, even if the complexity of the relationship between sensor data and occupancy rate would otherwise warrant it. It was therefore decided to start with a single hidden layer structure. As a starting point, the number of neurons was chosen to be around the same order of magnitude as the number of inputs for a single zone. Figure 6-8 shows the average Root Mean Square Error (RMSE) over all zones on the training and test data for varying numbers of neurons. RMSE is

used to show the magnitude of errors without allowing positive and negative errors to cancel out. It can be seen from the increased test RMSE that overfitting of the network is an issue even at the lower numbers of neurons, likely due to the small size of the training dataset and the relatively large number of sensor inputs. During initial tests, a structure with 10 hidden neurons was proposed: here, the model may be able to reflect some complexity of relationship, but avoids the more severe overfitting issues with higher numbers of neurons. As mentioned above, the optimum weights in the neural network can be found using a range of optimisation techniques. By default, the Matlab ANN toolbox uses Levenberg-Marquardt optimisation method [214] for the training process: this method was used during the following tests unless otherwise stated, with alternative methods explored in later sections.

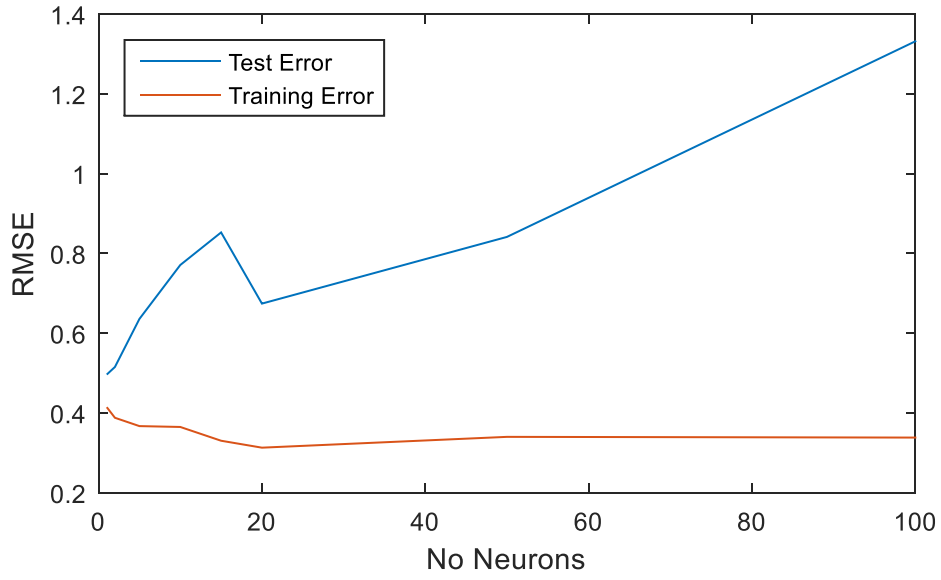


Figure 6-8 - Error on the Test and Training Data with Varying No. Neurons

The data used for the following model development is taken from two samples of manually-recorded occupancy data, one from the Phase 1 sensor arrangement, and one from the Phase 2 sensor arrangement as described in Section 5.2. The default Matlab ratio between training and test data was used: with 70% of available data used for training, 15% for validation and 15% for testing. Given the tendency for conditions in the building to stay relatively steady for multiple consecutive timesteps, it was found that the standard practice of selecting the test and validation data samples randomly from the whole dataset gave overly optimistic low errors on the test data results. This occurred because most of the test data samples had closely corresponding

entries with very similar sensor values within the training data, for which the model had been specifically optimised. This meant that overfitting to the training data was not properly indicated on the test data results, but had the model been run on a new week of sensor data, the model performance would be extremely poor. To avoid this issue, in the following work the training, test and validation data were sampled as continuous blocks from the whole dataset, with approximately the first five days of each test week assigned for training data and the last two days assigned for validation and testing respectively. With this continuous block sampling, the test data samples do not have direct equivalents in the training dataset, and so a more realistic idea can be gained on the model's overfitting issues/performance on previously unseen sensor data.

6.4 Tendency to find Local Minima

As Neural Networks typically optimise their parameters using some form of gradient descent, it is possible for networks with a complex cost function to get 'stuck' in local minima of the cost function without finding the global optimum values for the internal weights. This concept is illustrated in Figure 6-9, where a possible network cost function over all possible weight values is visualised as a 3D surface. As the network is trained, the weights are initially assigned random values and the weights are adjusted in steps towards a lower value of the cost function until a minimum is reached. If the random initialisation of the weights happens to fall close to point 1, the network is likely to find the minimum cost at point A, which in this case is the global minimum. However, if the random initialisation falls closer to point 2, the network may find the local minimum at point B, leading to a higher cost than the global minimum.

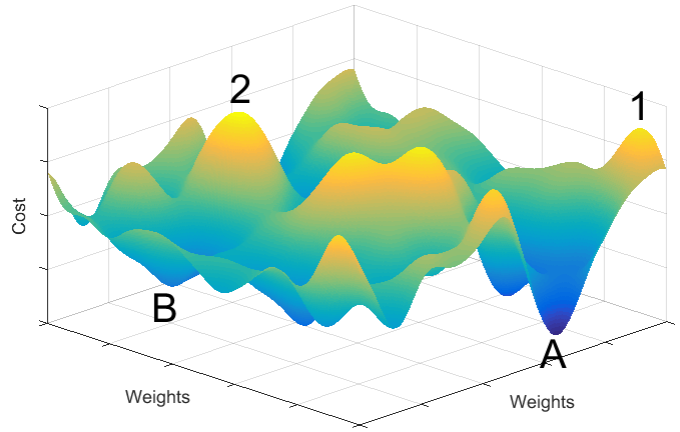


Figure 6-9 - Illustration of an ANN Cost Function with Local Minima

In many networks, the possibility of finding a local minimum is not considered a problem, as in many cases the cost function tends to be more valley-shaped than the illustration in the figure or local minima tend to take similar values. However, given the complex relationship between noisy sensor data and occupancy rates, and the planned methods of comparing error rates of various sensor configurations to optimise how data is presented to the network, it was decided to investigate whether local minima could be an issue in this study.

A network was trained for each zone using the full set of environmental sensors attributed to these zones. This method was repeated 100 times, with the outputs of the network recorded each time. Figure 6-10 shows the distribution of the error rates from these 100 trainings. It can be seen that the error rate of the same neural network structure can vary significantly depending on the randomised initialisation, suggesting that the network training is prone to falling into local minima and that this may be an issue for this particular application.

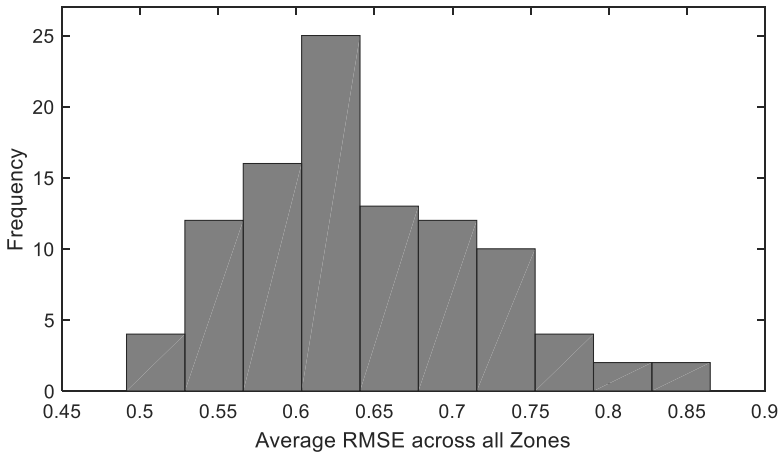


Figure 6-10 - Distribution of Root Mean Squared Error from 100 randomised-initialisation trainings of the same ANN structure and inputs

With several alternative training functions available in Matlab, it was investigated whether any were more effective than the default Levenberg-Marquardt optimisation, with the results shown in Table 6-1. It was found that the Levenberg-Marquardt optimisation, alongside Conjugate Gradient-based methods, gave a relatively low mean error with the lowest variation in outcomes. It was therefore decided to proceed with Levenberg-Marquardt training as a basis for comparing the performance of different network inputs. In order to reduce the effect of falling into local minima when comparing configurations, the RMSE used in the following tests was calculated by training the same structure at least 5 times and averaging RMSE values.

Table 6-1 – RMSE Mean and Range over 100 trainings of the same ANN structure for all default training functions available in Matlab

Training Function	Mean RMSE	RMSE Range
Levenberg-Marquardt	0.626	0.358
Conjugate Gradient with Powell/Beale Restarts	0.628	0.349
Polak-Ribiére Conjugate Gradient	0.633	0.298
Fletcher-Powell Conjugate Gradient	0.635	0.345
BFGS Quasi-Newton	0.650	0.882
Scaled Conjugate Gradient	0.661	0.429
One Step Secant	0.693	0.790
Resilient Backpropagation	0.754	0.522
Gradient Descent	0.783	0.995
Variable Learning Rate Gradient Descent	0.869	0.656
Bayesian Regularisation	1.118	2.480
Gradient Descent with Momentum	1.536	1.333

6.5 Model Optimisation

As found in the previous sections, an ANN model trained on all available sensors for a given zone tends to overfit, showing a significantly larger error on test data than the training data. Figure 6-11 illustrates this issue for a single zone: it can be seen that, in the test day, the occupancy rate is never estimated to be at zero, even during the night. As gathering a reliable training dataset of a longer timeframe was not viable with the data collection techniques used, it was necessary to explore options to address the overfit issue.

Model Combination/Mixture of Experts – One way to improve overfitting issues and the estimations made for inputs not in the training set is to average the estimations from several models. Deliberately training models that make varied, although equally accurate, predictions can allow an aggregate performance that is much more effective than any single model alone [215].

Regularisation – As the model is trained, the cost function is used to ensure that the overall error is reduced at each step. Assigning an additional penalty in the cost function for higher model weights can penalise overly complex models, creating a bias against including superfluous variables and tending towards a smoother fit to the training data.

Model Structure – One of the more effective ways to reduce overfitting issues is to reduce the number of input variables or features, discarding any that do not contribute meaningful information to the network. A potential starting point was to test the information gained from each of the sensor types.

It was decided that the initial model optimisation would be to remove any of the input features that were not providing useful information to the network. If overfitting issues were still present after this process, further options were to be explored.

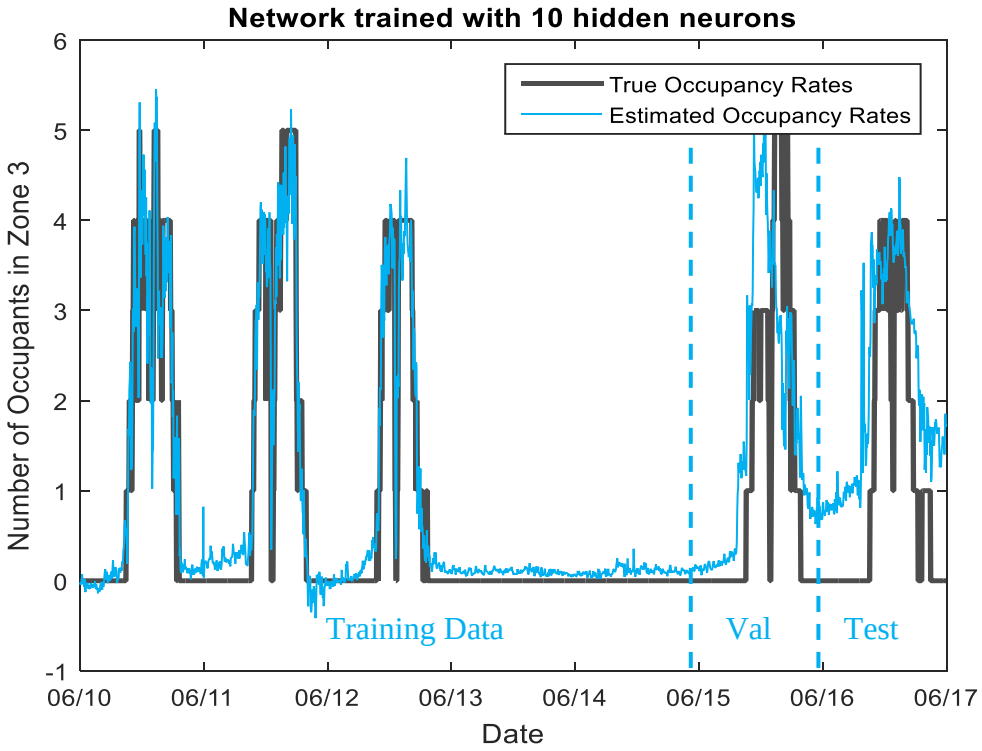


Figure 6-11 - ANN training, validation and test outputs for Multi-Occupant Office, Trained on all Zone Sensors

6.5.1 Single Sensors for Manual Feature Selection

One way to measure the information gain from each sensor was to train an ANN with data from a single sensor as the input. By comparing the error in the estimated number of people on the test data after training with each sensor, an idea of how relevant each sensor is to occupancy levels can be formed. It should be stressed that this is not a perfect approach to prioritising sensors, as some sensors may produce a high error alone but provide valuable context to the readings from other sensors (for example, the effect of opening a window on the CO_2 level). The comparison of single sensors is intended to be simply a starting point to systematically rank sensors where no other logical distinctions can be applied.

Phase 1 Data

Table 6-2 shows the RMSE for each sensor type when used as the sole input for an ANN trained on one week of manually labelled data, with a visual representation of each sensor's results in Figure 6-12. From this, it could be assumed that the window

use, ground floor PIR and CO_2 sensor provide the most occupancy-relevant information, and should be included in any reduced feature set for this zone.

Table 6-2 - Average Error in No People Estimated by ANN Trained on Single Sensors, Single Occupancy Office

Window	Ground Floor PIR	CO_2 Sensor	First Floor PIR	Temperature	Ext Door	Humidity
0.460	0.541	0.549	0.600	0.610	0.614	0.669

It should be noted that there are likely some false equivalencies, as in the test week it happens that the ground-floor single office occupant is present during the same hours as occupants triggering the upstairs PIR, for example. The external door sensor was also not used during the test week – meaning that it provided no information despite the similar window sensor being highly ranked. This is one of the disadvantages of attempting to train a comprehensive model on a single week of manually labelled data, and can only be guaranteed to be improved by collecting training data over a longer period. Unfortunately, due to the intrusive and time-consuming nature of collecting manual occupancy data from 10-15 occupants, it was not considered feasible to collect a larger training dataset in this manner.

Table 6-3 shows the sensors for all Phase 1 zones, ranked by individual error rate. Some general patterns were noted:

- Local CO_2 level was in the top 3 features for all but one zone – this confirms previous research findings on the value of CO_2 as an occupancy measure.
- The ground floor PIR ranked higher than the upper floor PIR for all zones – even the ones on the upper floor (Zone 1) – this could indicate a coincidental correlation due to the limited training data, or a potential fault with the upper floor PIR hardware or location. It should also be noted that the ground PIR was placed in an area that is passed when accessing most other spaces in the building, while the upper floor sensor is more isolated.
- Window behaviours were indicative of occupancy during Phase 1, which was conducted during a summer period. It should be noted that this pattern will not follow through to winter behaviours. If training a model on summer data, care must be taken to check if the model still works on winter data for the same zone.

Table 6-3 - Ranking of Information Gained from Single Sensors for each Zone – Phase 1

	Zone 1 (multi office)	**Zone 2** (meeting space)	**Zone 3** (multi office)	**Zone 4** (single office)	**Zone 5** (kitchen)	**Zone 6** (display space)	**Zone 7** (display space)	**Zone 8** (corridor)
1	CO_2	Velux 2	GF PIR	Window	CO_2	Temperature	CO_2	GF PIR
2	GF PIR	Velux 5	CO_2	GF PIR	Window 2	Ext Door	Humidity	UF PIR
3	UF PIR	Velux 3	UF PIR	CO_2	Window 1	CO_2	Temperature	
4	Temperature	Temperature	Window 4	UF PIR	GF PIR	Humidity	GF PIR	
5	Window Velux	CO_2	Window 2	Temperature	Temperature	GF PIR	UF PIR	
6	Window	Humidity	Window 3	Ext Door	Ext Door	UF PIR		
7	Humidity	Velux 4	Window 1	Humidity	UF PIR			
8	Ext Door	Velux 1	Temperature		Humidity			
9		GF PIR	Humidity					
10		UF PIR						
11		Ext Door						

Table 6-4 shows the RMSE of networks trained with reduced feature sets based on the individual sensor error rates as calculated above. A clear correlation can be seen between the number of features included and the error rate, with a reduced number of inputs giving a reduced error. This follows the hypothesis that the baseline ANN structure suffered from overfitting to the data. From the results in Table 6-4, it appears that 1-5 features is the optimum on average, providing a 28% reduction to the baseline average RMSE.

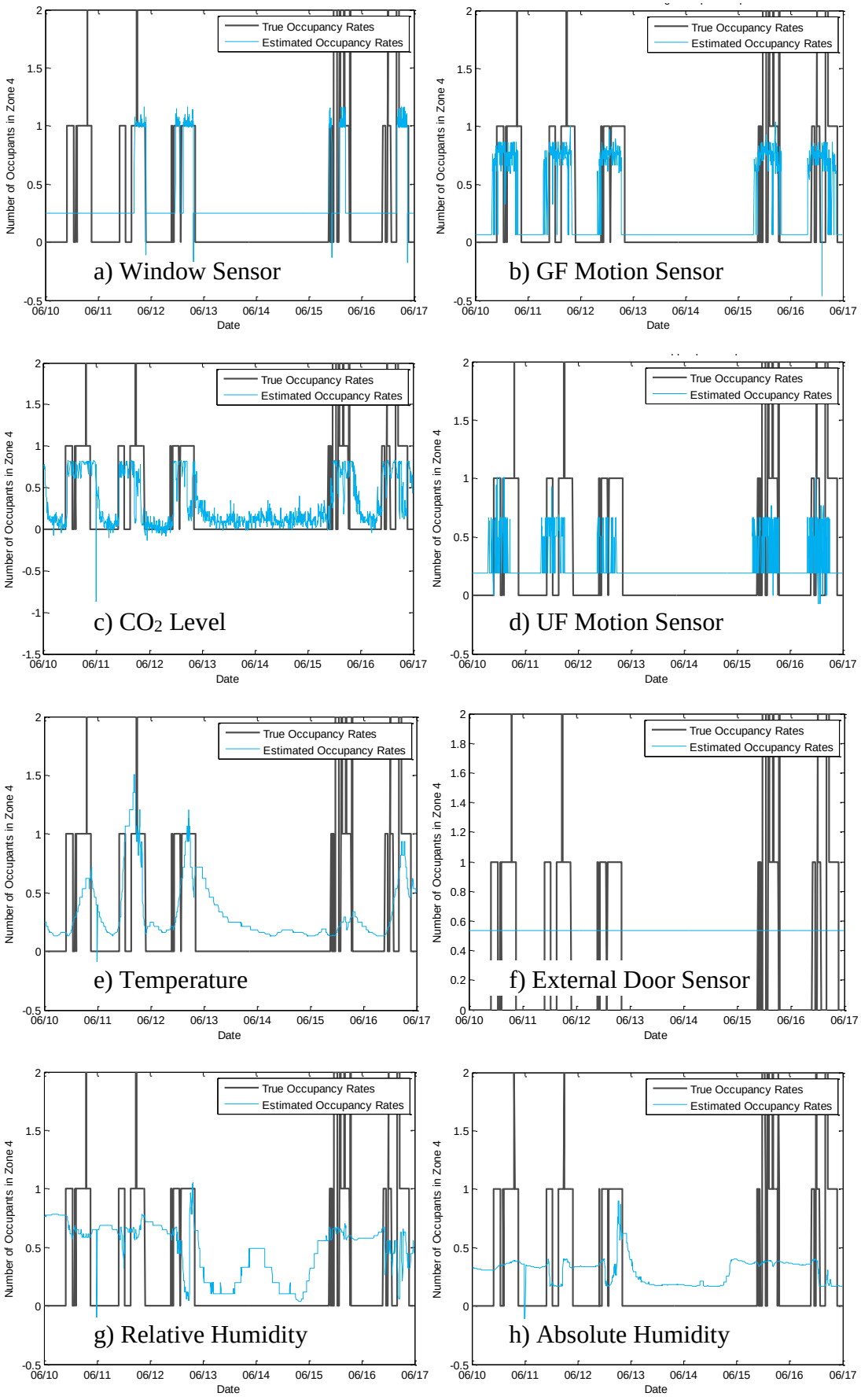


Figure 6-12 - Results of ANN Trained on Individual Sensors for Single-Occupancy Office

Table 6-4 – Manual Feature Selection RMSE against Baseline ANN Structure – Phase 1

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5	Zone 6	Zone 8	Avg*
Baseline	**1.135**	**1.089**	**1.166**	**0.596**	**0.319**	**0.118**	**0.401**	**0.689**
1 Feature	0.964	0.397	0.957	0.453	0.252	0.099	0.273	0.485
2 Features	0.929	0.432	0.888	0.421	0.446	0.099	0.305	0.503
3 Features	0.962	0.439	0.908	0.420	0.344	0.099		0.494
4 Features	0.961	0.532	0.905	0.419	0.318	0.100		0.503
5 Features	0.963	0.462	0.879	0.445	0.332	0.101		0.495
6 Features	0.989	0.694	0.936	0.430	0.501	0.102		0.562
7 Features	0.921	1.492	1.005	0.447	0.343			0.657
8 Features	1.000	1.011	0.936		0.332			0.588
9 Features		0.707	1.002					0.554
10 Features		1.309						0.640
11 Features		2.128						0.757

* Average calculated using the maximum no. features in zones with fewer than the stated number of features

Phase 2 Data

The above process was also conducted on the data from Test Phase 2 for comparison. In this data, more zones of the building are included in the study.

The ranked sensors for each zone are shown in Table 6-5. In comparison to the Phase 1 data rankings:

- The newly included PIR motion sensors ranked highly for information gain in all zones. This highlights the value of motion sensing for effective occupancy detection.
- The CO_2 level ranked similarly in several zones, but is notably placed last for information gain in five of the zones tested in Phase 2: the meeting space, kitchen, upper floor single office and basement level offices. For the basement office LG01, this can be explained by patchy data from the CO_2/temp/humidity sensor due to distance from the wireless signal receiver, which changed locations between Phase 1 and Phase 2. For the other zones, the CO_2 sees a decreased ranking in the winter test period versus the summer: this was counter to expectations, given the tendency for occupants to open windows and affect CO_2 concentration during the summer months. This may

indicate a negative interaction with the mechanical ventilation system, which was in operation during the winter month testing.

- One pattern to note is that the rooms with a low CO_2 ranking were occupied intermittently during the Phase 2 test period. This may indicate the time lag of increasing CO_2 levels can be counterproductive to a quick detection of occupants when the duration of occupancy is low.
- The window opening behaviour during this period continued to provide information gain in the meeting and kitchen spaces. In other zones such as the single offices A04 and A05, as was expected from the Phase 1 testing, the windows were not used as frequently during the heating period and so were not as indicative of occupancy.

Table 6-5 - Ranking of Information Gained from Single Sensors for each Zone – Phase 2

	'MGH A01' (meeting space)	**'MGH A02' (multi office)**	**'MGH A03' (kitchen)**	**'MGH A04' (single office)**	**'MGH A05' (single office)**	**'MGH B01' (single office)**	**'MGH B02' (multi office)**	**'MGH LG01' (multi office)**	**'MGH LG03' (display space)**
1	PIR	CO_2	PIR	PIR	PIR	PIR	PIR	Ext Door	Temperature
2	Velux 3	PIR	Window 1	Temperature	CO_2	Ext Door 2	CO_2	PIR	Humidity
3	Velux 4	Window 3	Window 2	Window 1	Ext Door	Ext Door 1	Window 1	Humidity	CO_2
4	Velux 1	Ext Door	Ext Door	Ext Door	Window 1	Window 1	Velux 1	Temperature	
5	Velux 2	Window 1	Humidity	CO_2	Temperature	Humidity	Humidity	CO_2	
6	Velux 5	Temperature	Temperature	Humidity	Humidity	Temperature	Temperature		
7	Ext Door	Window 2	CO_2			CO_2			
8	Temperature	Humidity							
9	Humidity								
10	CO_2								

Table 6-6 shows the performance of reduced feature sets relative to a baseline case using all sensors available for each zone. As with the Phase 1 data, features were chosen in order of information gain as shown in Table 6-5, estimated from the error rate when trained on each sensor input alone. It can be seen that a reduced sensor set gives a lower average error, as with Phase 1. However, the highest error rate is seen at 4 features. This is due to the jump in error rate on office LG01, highlighted in the

table. When this office is omitted, the error rate is similar for the first 5 features in the other zones, with an increase in error after 5 features. The RMSE with 5 features shows a reduction of 14% from the full feature set: a smaller reduction seen than in Phase 1 testing. This suggests a possibility that the overfitting problem may not be as present with the Phase 2 data.

Table 6-6 – Manual Feature Selection RMSE against Baseline ANN Structure – Phase 2

	'MGH A01' (meeting space)	'MGH A02' (multi office)	'MGH A03' (kitchen)	'MGH A04' (single office)	'MGH A05' (single office)	'MGH B01' (single office)	'MGH B02' (multi office)	'MGH LG01' (multi office)	'MGH LG03' (display space)	Avg*	Avg* w/o LG01
Baseline	**0.76**	**1.02**	**0.17**	**0.38**	**0.21**	**0.38**	**0.63**	**1.84**	**0.00**	**0.60**	**0.45**
1 Feature	0.49	0.85	0.16	0.39	0.17	0.28	0.70	0.49	0.00	0.39	0.38
2 Features	0.49	0.85	0.16	0.61	0.18	0.28	0.63	0.60	0.00	0.42	0.40
3 Features	0.49	0.85	0.16	0.70	0.18	0.28	0.63	1.85	0.00	0.57	0.41
4 Features	0.49	0.86	0.16	0.49	0.18	0.28	0.63	2.20		0.59	0.39
5 Features	0.49	0.85	0.18	0.49	0.18	0.30	0.61	1.54		0.52	0.39
6 Features	0.49	1.27	0.17	0.44	0.22	0.32	0.60			0.56	0.44
7 Features	0.50	1.11	0.18			0.32				0.55	0.42
8 Features	0.59	0.97								0.54	0.41
9 Features	0.54									0.53	0.41
10 Features	0.71									0.55	0.43

* Average calculated using the maximum no. features in zones with fewer than the stated number of features

It should also be noted that the RMSE of zones included in Phase 1 and Phase 2 shows a reduction of error with the inclusion of the PIR sensors in Phase 2, as shown in Table 6-7. Some change in the error rate can be attributed to changed space uses: in particular the zones B02, which changed office users between tests, and LG01, which was converted to a shared office space between tests and so saw a much higher occupancy during Phase 2. However, the reduced error rate on other zones suggests the value of including local PIR data to correctly detect occupancy rates.

Table 6-7 – Average RMSE of full feature set – Phase 1 vs Phase 2

	'MGH B01'	'MGH A01'	'MGH A02'	'MGH A05'	'MGH A03'	'MGH LG01'	'MGH LG03'	Avg
RMSE Phase 1	1.135	1.089	1.166	0.596	0.319	0.118	0.401	0.689
RMSE Phase 2	0.629	0.761	1.022	0.215	0.171	1.845	0.000	0.663

6.5.2 Sensor Combinations

As mentioned in the previous section, it is possible that some sensors may be of more value when combined with another sensor type than they are alone. Where two measured variables interact, such as the opening of a window affecting the CO_2 level, it is logical that a network trained on both of these variables together should be able to gain more information. For the single-occupant office tested in the previous section, three logical sensor combinations were tested:

- CO_2 level and window opening,
- Relative humidity and window opening,
- Absolute humidity, calculated from RH and temperature.

The relative humidity (Figure 6-12 g) was surprisingly low in terms of information value. It was investigated whether absolute humidity (Figure 6-12 h) would provide more information after accounting for temperature changes. While this did negate the peak in estimated occupancy seen at the weekend in the graph and improve the average error, it can be seen that there is still little to no sign of changes in occupancy during weekdays, suggesting that the absolute humidity is still not a valuable occupancy measure.

For the other sensor combinations summarised in Table 6-8, the combination of sensors appears to improve the error rate more significantly. In particular it should be noted that the combination of CO_2 level and window opening gave a lower error rate than either sensor alone for the Phase 1 testing.

Table 6-8 - Average Error in No People Estimated by ANN Trained on Single/Combined Sensors, Single Occupancy Office, Phase 1 Data

CO_2 & Window	Window	RH% & Window	Ground Floor PIR	CO_2	Temp	Abs Humidity	RH%
0.427	0.454	0.489	0.542	0.546	0.572	0.610	0.823

While these manually constructed sensor combinations in general showed a positive effect in the Phase 1 testing during warm weather conditions, the same windows were found not to be opened at all during the Phase 2 testing in cold weather conditions. This highlights some of the issues around models that rely on short periods of training data: a model trained during the summer may rely heavily on window-related behaviours that are not present during winter, meaning that the model performance

would be significantly reduced at different times of year, or when users with different habits move into the same space. To avoid these problems, the model would need to be trained on deliberately diverse training data, or periodically updated. Due to the change in sensor layout between Phases 1 and 2 of this study, it was not possible to train a model on both summer and winter data at once in this work.

6.5.3 Principal Component Analysis (PCA)

Aside from manually selecting a reduced set of inputs to the ANN, the set of features can also be reduced using Principal Component Analysis (PCA). This technique can be used where the set of input variables have some correlation to each other: the data is mapped to a new set of variables based along the axis of greatest variation in the original data [216]. In this way, the technique aims to represent the greatest variation from a dataset within a reduced number of variables, allowing the total number of variables to be reduced.

In Matlab, a function is available to calculate the principal components of a dataset using singular value decomposition [217]. The data must first be normalised, to ensure that the numerical value of data (for example, CO_2 data ranged between 500-1000ppm and Wi-Fi presence ranged between 0-1) does not skew the results towards the variables with the greatest numerical variance. Normalising ensured that all inputs were rescaled to range from -1 to 1.

6.5.3.1 Phase 1 Data

Table 6-9 shows an example of the calculated principal components for the single-occupant office zone. By default, the components are ordered from most to least variation, meaning that, in theory, the most useful information should be encoded in the first few components. The first principal component features several sensor types with near-equal weighting, suggesting that the CO_2, temperature, window use and motion sensors tend to vary together, with a positive correlation. The second PC features the humidity and temperature, while the third features the window most strongly. The final PC, which describes the least variance, features only the external door, which did not see any use during the tested period, and so reported a constant value.

Table 6-9 - Principal Component Coefficients for Single-Occupant Office

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Zone 4 CO_2 Level	0.51	-0.06	-0.34	-0.48	0.19	0.60	0.00
Zone 4 Humidity	-0.12	0.64	-0.50	0.46	0.30	0.16	0.00
Zone 4 Ext Door	0.00	0.00	0.00	0.00	0.00	0.00	1.00
Zone 4 Temp	0.39	0.56	-0.10	-0.33	-0.46	-0.45	0.00
Zone 4 Window	0.29	0.38	0.76	0.21	0.02	0.38	0.00
GF PIR	0.55	-0.16	0.05	0.19	0.61	-0.50	0.00
UF PIR	0.43	-0.32	-0.23	0.60	-0.53	0.13	0.00

One of the disadvantages of using PCA to reduce the number of model inputs is that PCA cannot identify where features vary due to occupancy independent factors. For example, the internal temperature can see a large variation depending on the weather and time of day, but may not be greatly affected by occupancy rates. As PCA seeks only to maximise the variance in each component, a varied temperature is likely to be given more importance than necessary for occupancy detection. This drawback may be the cause of the unstable effect on error rate on networks trained on different numbers of PCs, as seen in Table 6-10. It appears from this table that the number of PCs used as ANN inputs does not have a simple correlation with the average RMSE. However, it can be seen that the RMSE is reduced for some numbers of PCA components relative to the baseline ANN structure. 4-5 principal components appears to be the optimum on average giving a 25% reduction on the baseline error, although the unstable variation suggests this number may be coincidental.

Table 6-10 - PCA RMSE in Comparison to the Baseline ANN – Phase 1

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5	Zone 6	Zone 8	Avg*
Baseline	**1.135**	**1.089**	**1.166**	**0.596**	**0.319**	**0.118**	**0.401**	**0.689**
1 PC	1.039	1.299	0.990	0.470	0.281	0.098	0.268	0.635
2 PCs	0.881	0.720	0.928	0.456	0.301	0.099	0.304	0.527
3 PCs	0.883	1.094	0.859	0.434	0.317	0.099		0.570
4 PCs	0.891	0.725	0.875	0.439	0.305	0.100		0.520
5 PCs	0.971	0.677	0.801	0.437	0.308	0.099		0.514
6 PCs	0.950	1.031	0.986	0.417	0.279	0.100		0.581
7 PCs	0.991	1.186	1.051	0.415	0.288			0.619
8 PCs	0.952	1.819	0.927		0.293			0.687
9 PCs		0.666	0.961					0.527
10 PCs		1.134						0.594
11 PCs		0.805						0.547

* Average calculated using the maximum no. PCs in zones with fewer than the stated number of features

6.5.3.2 Phase 2 Data

Table 6-11 shows the average RMSE of networks trained with different numbers of PCs for the Phase 2 data. It can be seen that the tendency for fewer PCs to give a lower rate of error is somewhat reversed in the Phase 2 test, with lower error rates as more features were included. This is somewhat counter to the expectation that reducing the number of inputs would decrease error rates from overfitting. On inspection of the PCs, it was found that the less useful environmental measures that tend to change together such as the humidity and temperature had been ‘bundled’ into the first few PCs, with the more valuable motion sensor data featured more highly in the later PCs. The percentage reduction in error that could be achieved with a reduced number of ANN inputs was similar to the Phase 1 data, with a maximum reduction of 27% at 5 features.

Table 6-11 - PCA RMSE in Comparison to the Baseline ANN – Phase 2

	'MGH A01' (meeting space)	'MGH A02' (multi office)	'MGH A03' (kitchen)	'MGH A04' (single office)	'MGH A05' (single office)	'MGH B01' (single office)	'MGH B02' (multi office)	'MGH LG01' (multi office)	'MGH LG03' (display space)	Avg*
Baseline	**0.761**	**1.022**	**0.171**	**0.381**	**0.215**	**0.380**	**0.629**	**1.845**	**0.000**	**0.601**
1 PC	0.602	1.075	0.161	0.515	0.455	0.305	0.974	2.268	0.000	0.706
2 PCs	0.675	0.918	0.164	0.383	0.335	0.295	0.574	1.943	0.000	0.587
3 PCs	0.582	0.867	0.164	0.428	0.198	0.309	0.584	1.089	0.000	0.469
4 PCs	0.495	1.219	0.163	0.396	0.188	0.373	0.568	1.303		0.523
5 PCs	0.462	0.961	0.170	0.350	0.193	0.316	0.585	0.918		0.440
6 PCs	0.586	0.819	0.163	0.356	0.199	0.362	0.557			0.440
7 PCs	0.841	0.836	0.162			0.346				0.468
8 PCs	0.634	0.848								0.447
9 PCs	0.626									0.446
10 PCs	0.617									0.445

* Average calculated using the maximum no. PCs in zones with fewer than the stated number of features

6.5.4 Signal smoothing and pre-processing

One of the limitations identified with using only the latest raw sensor data is the lack of any trend or temporal value to the data inputs. As each 5-minute step of the detection model takes only a single data value per sensor, it is also highly susceptible to noise on the sensor inputs.

Two potential solutions to this issue were identified: smoothing the data from noisy sensors using a moving average as input rather than the raw data values, or pre-processing the data to identify trends in the data as they occur. Processing trends in the sensor data was identified as a potentially valuable source of extra information, especially for air quality measures such as CO_2 concentration, for which the absolute value is less indicative of occupancy than whether the concentration is rising or falling. For the following tests, CO_2 concentration data was used to test the effectiveness of each strategy.

Figure 6-13 shows a sample of raw CO_2 data against two datasets smoothed using moving averages. While it can be seen that the noise on the data is significantly reduced while the CO_2 level is steady during the night, the moving average introduces

a lag on peaks/troughs. In the example highlighted in the figure, the transition from falling to rising is delayed by 30 min for the simple moving average and 15 min for the exponential moving average. In a control system where changes in occupancy need to be detected as quickly as possible, introducing further time lag to an already slow-response data source may prove to be more of a hindrance than a help.

Figure 6-14 and Figure 6-15 show the pre-processed CO_2 variable against a sample of raw CO_2 data. In this case, the variable used was the gradient of a linear equation fitted to the previous hour's data points. In practice, this could be calculated in real time using the Matlab 'polyfit' function, and should have a faster response to sudden changes in CO_2 concentration than the moving average. The gradient provides a clearer distinction between when the CO_2 is rising or falling than the raw data alone. However, the first figure shows that the gradient still includes some amount of noise, with nonzero gradients present throughout the unoccupied night-time periods. A second, filtered version of the gradient variable was introduced, which set all gradients below a threshold value to 0.

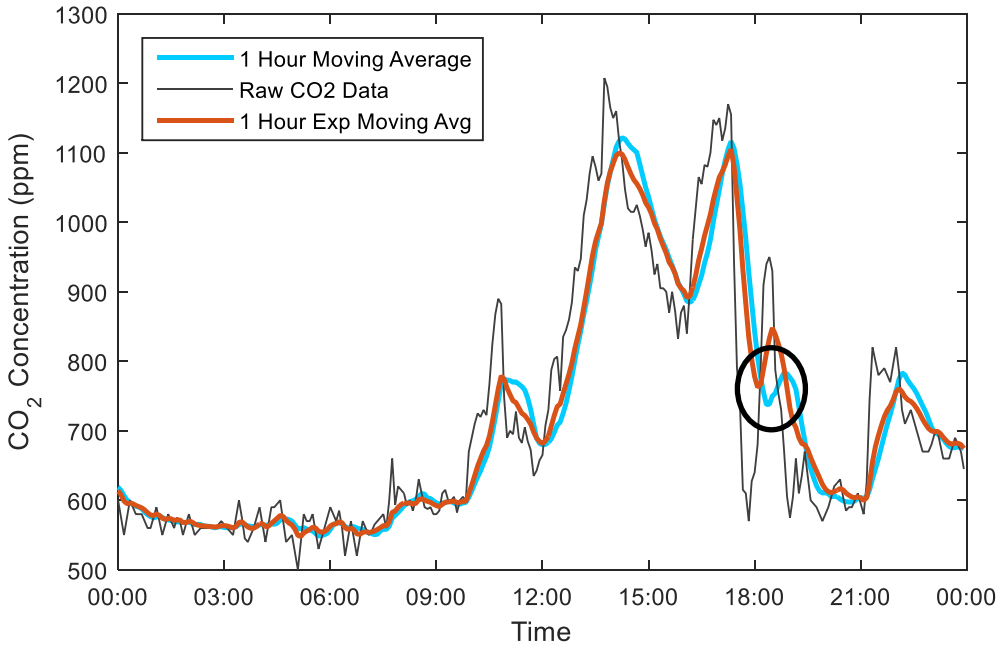


Figure 6-13 - Comparison of Moving Average Smoothed CO_2 Data

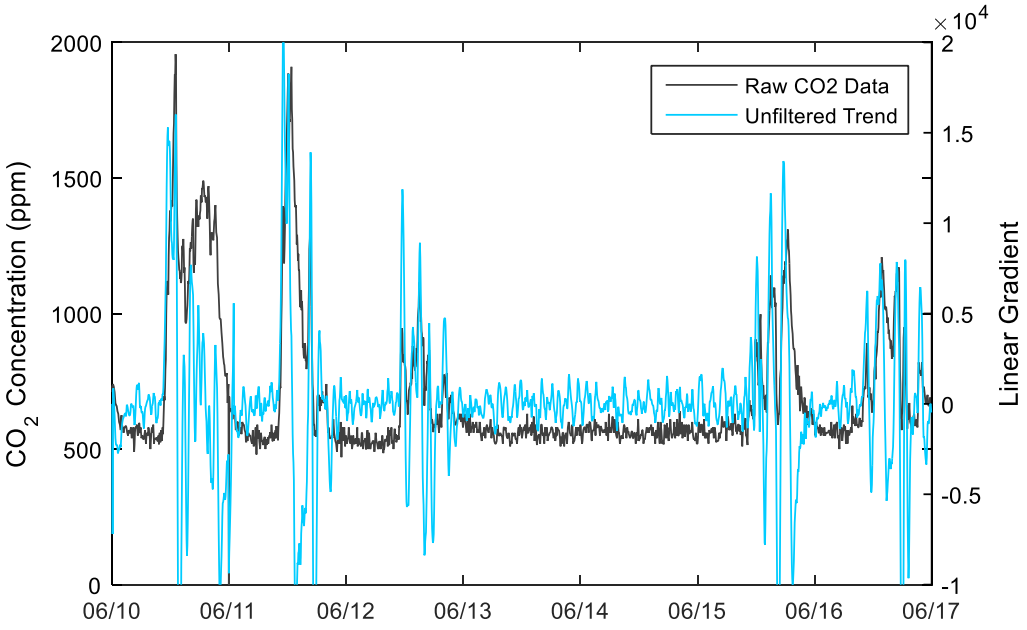


Figure 6-14 – Comparison of CO_2 Trend Variable and Raw CO_2 Data

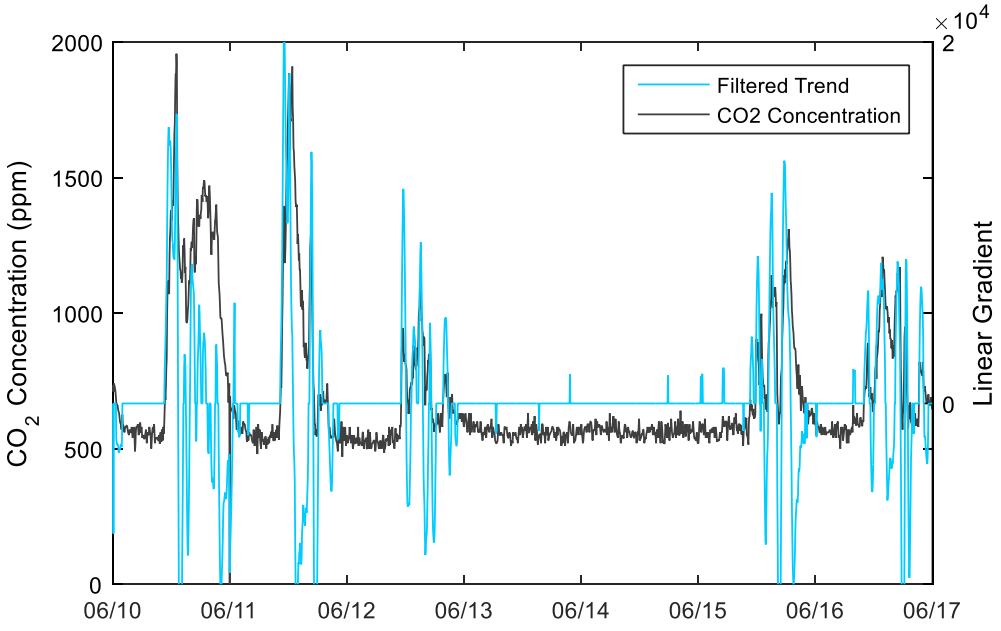


Figure 6-15 - Comparison of Filtered CO_2 Trend Variable and Raw CO_2 Data

All of the above processed variables were tested for usefulness by training a room-level ANN, using Levenberg-Marquardt training, for each monitored zone of the building, omitting zones without CO_2 data and without occupancy during the test period (Zones 7 and 8). Each zone was given a designated list of relevant sensors, plus some combination of the processed variables relevant to the zone. A summary of the average error over 5 randomised-initialisation ANN trainings of various combinations of the processed data is shown in Table 6-12 and Figure 6-16. Using the

RMSE from a network trained on the unprocessed data only as a baseline, it can be seen that both moving average variables increased the RMSE, likely due to the increased lag highlighted in Figure 6-13. The average RMSE is not significantly affected when replacing CO_2 with linear gradient trend data. The RMSE is reduced when the network is trained with a combination of raw CO_2 and trend data. It is therefore concluded that, on average, supplying a pre-processed linear gradient with the raw CO_2 data is the most effective combination. It should be noted that the decrease in RMSE is not consistent across all zones: with no significant decrease to RMSE in the multi-occupant offices, or the single-occupant office. The greatest decrease in error was seen in the intermittently used spaces, in particular the meeting room. This may suggest that the CO_2 trend data is most useful for detecting occupancy events that do not last long enough to reliably push the CO_2 level above a threshold value.

Table 6-12 - Summary of Average RMSE over 5 trainings of ANN with pre-processed CO_2 data – Phase 1 Data

	Zone 1 (multi office)	Zone 2 (meeting space)	Zone 3 (multi office)	Zone 4 (single office)	Zone 5 (kitchen)	Zone 6 (display space)	Average
Baseline CO_2	0.98	1.77	1.26	0.43	0.58	0.12	0.85
Trend, no filter	1.19	1.54	1.06	0.40	0.70	0.15	0.84
Trend, filtered	1.01	1.59	1.13	0.43	0.89	0.12	0.86
CO_2 and trend	0.99	1.23	1.11	0.45	0.52	0.14	0.74
CO_2 and trend, filtered	0.95	1.08	1.30	0.41	0.51	0.15	0.73
Simple moving avg CO_2	1.06	1.46	1.45	0.45	0.60	0.11	0.85
Exp moving avg CO_2	1.13	1.60	1.56	0.41	0.60	0.14	0.91

When compared in the same manner as the manual feature selection in the previous sections, the information gain from the CO_2 level and trend values can be assessed. Table 6-13 shows the ranking of these sensors for each zone tested in the Phase 1 period, where the trend data was ranked as more useful than the static CO_2 level in some spaces, in particular the intermittently occupied meeting space. This highlights the usefulness of observing whether the CO_2 level is rising or falling in zones that are not occupied for long enough to build up CO_2 levels beyond a predefined threshold. However, in summer months it can be seen that the CO_2 trend is not always reliable, a likely cause for the low ranking of trend in the kitchen area, where windows and

doors were opened during most occupied times, reducing the CO_2 level increase. During the heating season test in Phase 2 (Table 6-14) it can be seen that the CO_2 trend more consistently outranks the static CO_2 level. Once again this difference is particularly strong in the meeting space.

Table 6-13 - Compared Input Ranking of Raw CO_2 Data vs CO_2 Trend Gradient - Phase 1 Data

	Zone 1 (multi office)	Zone 2 (meeting space)	Zone 3 (multi office)	Zone 4 (single office)	Zone 5 (kitchen)	Zone 6 (display space)
CO_2 Ranking	1	6	3	4	1	4
CO_2 Trend Ranking	4	1	6	2	6	2

Table 6-14 - Compared Input Ranking of Raw CO_2 Data vs CO_2 Trend Gradient - Phase 2 Data

	'MGH A01'	'MGH A02'	'MGH A03'	'MGH A04'	'MGH A05'	'MGH B01'	'MGH B02'	'MGH LG01'	'MGH LG03'
CO_2 Ranking	11	2	7	6	3	8	2	6	3
CO_2 Trend Ranking	2	1	8	2	2	7	3	3	4

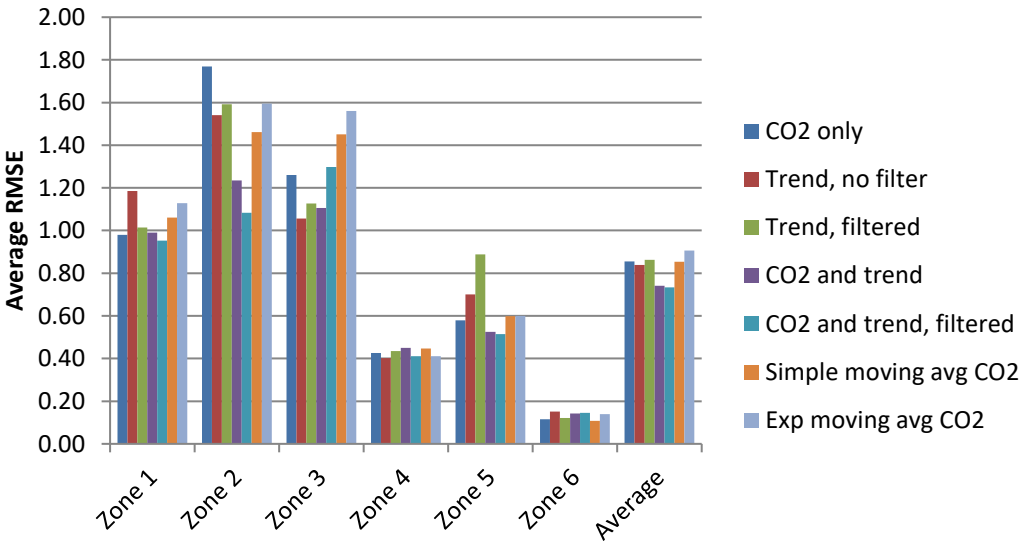


Figure 6-16- Summary of Average RMSE over 5 trainings of ANN with pre-processed CO_2 data

6.5.5 Wi-Fi Detection Data

In addition to the environmental sensors in the test building, data on the number of Wi-Fi-enabled devices present in the building was also collected during the testing periods. The method of collection and pre-processing of the Wi-Fi detection data is

discussed in detail in section 5.4.2. Data was received from the sensor as a series of discrete detection events, each with an associated device ID and signal strength. There are a variety of ways that this data can be included as an input to the ANN model. A range of these were tested in order to find the most effective way to utilise the Wi-Fi data:

- All data as a 0/1 presence for each discovered device: this is not feasible in the long term. For one, the inputs for the ANN model need to be constant, so any new devices would have no way to be included. This configuration was therefore not tested.
- Total number of devices seen: a sum of the number of devices seen at any given time.
- Total number of devices seen at each band of signal strength: summed number of devices, separated by signal strength.
- Known individuals detected: from a predefined list of known building occupant device IDs, a 0/1 indicator for each occupant at each time step.
- Known individuals detected, signal strength: from a predefined list of known building occupant device IDs, a value indicator for each occupant at each time step showing the average signal strength received over the last 5 minutes.
- Combinations of the total and individual data described above.

As with the testing of CO_2 trend data, each of the Wi-Fi data types was tested against a baseline model for each zone, trained on the full set of environmental sensors associated with this zone. For each data type, an ANN model was trained using the same structure as the baseline. Table 6-15 shows the average error over 5 randomised-initialisation trainings for each set of data inputs during the Phase 1 test week. It can be seen that the average error rate is decreased with the inclusion of the total number of devices seen, while the error is increased when only the presence of known individuals is included. As each individual is represented to the network as a separate input feature, this likely indicates that the individual data from this period is not providing enough worth to counter the increased tendency for the network to overfit when supplied with more features.

Table 6-15 - Average RMSE per Zone for ANNs trained with different Wi-Fi Data Inputs – Phase 1 Data

	Baseline – no Wi-Fi	Wi-Fi Total Only	Totals by Signal Strength	Individuals 0/1	Individuals Signal Strength	Total and Individuals	Totals and Individuals by Strength
Zone 1 (multi office)	**1.01**	0.93	0.92	1.02	1.12	1.03	0.98
Zone 2 (meeting space)	**2.31**	1.02	1.38	1.31	1.21	1.10	1.12
Zone 3 (multi office)	**0.96**	0.88	0.92	1.50	1.38	1.37	1.43
Zone 4 (single office)	**0.42**	0.40	0.43	0.46	0.42	0.44	0.45
Zone 5 (kitchen)	**0.32**	0.32	0.31	0.55	0.56	0.64	0.67
Zone 6 (display space)	**0.10**	0.14	0.10	0.14	0.11	0.15	0.19
Zone 7 (display space)	**0.00**	0.00	0.00	0.00	0.00	0.00	0.00
Zone 8 (corridor)	**0.29**	0.31	0.29	0.62	0.53	0.47	0.39
Mean	**0.68**	**0.50**	**0.54**	**0.70**	**0.67**	**0.65**	**0.65**

This process was repeated for the data from the Phase 2 test week, with the results shown in Table 6-16. Once again, it can be seen that the inclusion of the total number of devices seen reduces the average error rate across all zones in the building. However, in the Phase 2 testing, the lowest error rates were achieved when the data for known individuals was included: a definite contrast to the Phase 1 tests. The largest error reduction from the baseline was seen in the multi-occupant office LG01 due to the poor environmental sensor coverage in this area, while the other zones show an error increase more consistent with the Phase 1 findings. This shows that the set of sensors that is most useful can be highly dependent on the specific circumstance of a zone's use.

Table 6-16 - Average RMSE per Zone for ANNs trained with different Wi-Fi Data Inputs – Average Results from 100 Trainings with Phase 2 Data

	Baseline – no Wi-Fi	Wi-Fi Total Only	Totals by Signal Strength	Individuals 0/1	Individuals Signal Strength	Total and Individuals	Totals and Individuals by Strength
MGH A01 (meeting space)	0.780	0.765	0.771	0.554	0.577	0.584	0.624
'MGH A02' (multi office)	1.035	1.090	0.991	0.943	0.937	0.988	1.000
'MGH A03' (kitchen)	0.185	0.186	0.220	0.211	0.222	0.219	0.221
'MGH A04' (single office)	0.410	0.416	0.467	0.487	0.480	0.506	0.515
'MGH A05' (single office)	0.218	0.213	0.240	0.261	0.242	0.262	0.292
'MGH B01' (single office)	0.361	0.330	0.353	0.395	0.377	0.385	0.382
'MGH B02' (multi office)	0.608	0.569	0.631	0.670	0.681	0.654	0.690
'MGH LG01' (multi office)	1.936	1.679	1.181	0.675	0.963	0.753	0.831
MGH LG03' (display space)	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Other	0.296	0.300	0.305	0.304	0.306	0.304	0.308
Mean	**0.692**	**0.656**	**0.607**	**0.524**	**0.560**	**0.544**	**0.569**

The information gain of the Wi-Fi data relative to the environmental sensors was also tested using the same method as the feature selection: training networks with each of the sensor types as a single input and comparing error rates. For each monitored zone, the top ten sensors in order of information gain are presented in Table 6-17 and Table 6-18 for Phase 1 and Phase 2 respectively. Phase 1 shows that the total number of Wi-Fi devices detected over the whole house is a consistently information-rich feature for each of the zones. This input was valuable to the multi-occupant spaces in particular, such as the Research Office, PhD Office and Circulation Spaces in the building. As this feature is a global measure of the occupancy across the whole building, this association is logical. However, in the Phase 2 testing, the total number of devices was less highly ranked: this may have been due to the local PIR sensors fulfilling a similar role in providing an indicator for whether the space was occupied or not.

It should be noted that the Wi-Fi-detected presence of personal devices is a high-ranking feature for each of the zones. For some zones, the individual selected has a

logical association with the zone, as is the case with the single-occupant office, for which the highest rated feature is the presence of the office owner's mobile device. In many cases, however, there does not appear to be a logical connection between the individual and the zone – these cases are highlighted in the tables. These cases are likely due to coincidental overlap of the individual's presence and occupancy in a zone – a situation that could be better avoided with a larger training dataset. Given a limited set of training data, these features appear to be too prone to coincidental mis-assignment to provide benefit.

Table 6-17 – Wi-Fi Inclusive Ranking of Information Gained from Single Sensors for each Zone – Phase 1

	Zone 1 (meeting space)	Zone 2 (meeting space)	Zone 3 (multi office)	Zone 4 (single office)	Zone 5 (kitchen)	Zone 6 (display space)	Zone 7 (display space)	Zone 8 (Corridors)
1	Total Wifi Devices	Sunspace Window 5	Total Wifi Devices	Zone 4 Occupant	Other Occupant	Other Occupant	CO_2 Level	Total Wifi Devices
2	CO_2 Level	Zone 1 Occupant	GF PIR	Window 1	Total Wifi Devices	Zone 3 Occupant	Humidity	Zone 1 Occupant
3	Other Occupant	Temperature	UF PIR	Total Wifi Devices	CO_2 Level	Temperature	Temperature	GF PIR
4	GF PIR	Sunspace Window 2	CO_2 Level	GF PIR	Zone 3 Occupant	Zone 3 Occupant	GF PIR	Zone 3 Occupant
5	Zone 1 Occupant	Zone 3 Occupant	Other Occupant	CO_2 Level	Zone 3 Occupant	Zone 4 Occupant	UF PIR	Zone 3 Occupant
6	UF PIR	Zone 1 Occupant	Zone 3 Occupant	Temperature	Kitchen Window 1	Zone 1 Occupant	Total Wifi Devices	Zone 3 Occupant
7	Zone 3 Occupant	Sunspace Window 3	Zone 1 Occupant	Other Occupant	Kitchen Window 2	GF PIR	Zone 3 Occupant	Zone 1 Occupant
8	Zone 4 Occupant	Humidity	Zone 3 Occupant	Zone 3 Occupant	GF PIR	Zone 3 Occupant	Other Occupant	Zone 1 Occupant
9	Zone 3 Occupant	CO_2 Level	Zone 4 Occupant	UF PIR	Other Occupant	Basement Door	Zone 4 Occupant	Zone 4 Occupant
10	Humidity	Zone 3 Occupant	PhD Window 4	Zone 1 Occupant	Zone 4 Occupant	Zone 1 Occupant	Zone 3 Occupant	Other Occupant

Table 6-18 – Wi-Fi Inclusive Ranking of Information Gained from Single Sensors for each Zone – Phase 2

	MGH A01 (meeting space)	'MGH A02' (multi office)	'MGH A03' (kitchen)	'MGH A04' (single office)	'MGH A05' (single office)	'MGH B01' (single office)	'MGH B02' (multi office)	'MGH LG01' (multi office)	MGH LG03' (display space)	Other
1	PIR	A02 Occupant	B01 Occupant	PIR	PIR	A02 Occupant	PIR	A02 Occupant	Temperature	B02 Occupant
2	A02 Occupant	CO_2	A02 Occupant	A02 Occupant	A02 Occupant	PIR	Wifi Total	B01 Occupant	Humidity	A04 Occupant
3	B02 Occupant	B01 Occupant	B02 Occupant	B01 Occupant	A04 Occupant	A02 Occupant	CO_2	A02 Occupant	CO_2	LG01 Occupant
4	B01 Occupant	A04 Occupant	A04 Occupant	LG01 Occupant	B02 Occupant	A04 Occupant	B02 Occupant	LG01 Occupant	Wifi Total	A02 Occupant
5	A04 Occupant	A02 Occupant	A02 Occupant	B02 Occupant	A05 Occupant	A02 Occupant	A05 Occupant	A02 Occupant	A02 Occupant	LG01 Occupant
6	A05 Occupant	B02 Occupant	PIR	A02 Occupant	CO_2	A02 Occupant	LG01 Occupant	CO_2	LG01 Occupant	B01 Occupant
7	B02 Occupant	LG01 Occupant	A05 Occupant	A02 Occupant	Wifi Total	Ext Door 1	A02 Occupant	B02 Occupant	A02 Occupant	LG01 Occupant
8	LG01 Occupant	A02 Occupant	B02 Occupant	Temperature	A02 Occupant	Ext Door 2	B02 Occupant	A02 Occupant	B02 Occupant	A02 Occupant
9	A02 Occupant	PIR	LG01 Occupant	Window 1	B02 Occupant	LG01 Occupant	B01 Occupant	LG01 Occupant	LG01 Occupant	A02 Occupant
10	Window 1	A02 Occupant	Window 1	A04 Occupant	B01 Occupant	LG01 Occupant	A04 Occupant	LG01 Occupant	LG01 Occupant	A02 Occupant

In general, the above tests suggest that the total number of Wi-Fi enabled devices detected in a building can provide information useful to estimating local occupancy rates. Providing data on the signal strength of these detected devices showed mixed results on whether it provided more benefit than a global measure of devices detected. The presence of individuals can provide benefit depending on the specific circumstances of a zone's space use and occupants, but is also more likely to exacerbate overfitting issues with models trained on a limited training dataset. For the purpose of this study, where only short training datasets were available, it was decided to omit individual occupants from the final model. However, in cases where a longer period can be used for training, the problems with false association of individual devices should be reduced and so it would be recommended to include the binary presence of known individuals.

6.5.6 Bluetooth Beacon Data

During the Phase 2 testing, data was also collected from known occupants using personal mobile devices to detect local Bluetooth Beacons. The setup and analysis of this data collection method is presented in Section 5.4. As with the Wi-Fi data, there are multiple ways that this Bluetooth data could be represented to a NN model:

- Total number of occupants detected within the whole building: as the binary presence of occupants was shown to be more reliable than the exact location in initial testing (Section 5.4.5).
- Binary presence over whole building for each individual occupant: as above, but disaggregated to the level of each occupant.
- Total number of occupants logged in to each room over time, supplied as an integer value. It was expected that this would be of high value if the Bluetooth system was more reliable, but may not provide the best value given the known issues with exact location.
- Binary presence of each known occupant is supplied as an input to their 'main zone': this requires a manual assignment of a main zone for the occupants, based on the location of their working office desk.

The averaged results of 100 randomised initialisation trainings of each of these configurations are shown in Table 6-19. In these trials, the average error across the building was slightly decreased when the Bluetooth data was included in most formats, although the configuration with the greatest decrease was when occupants' data were manually assigned to their most frequently occupied zone. This is not ideal, as it requires further manual input and makes the system less robust to unusual occupant behaviours, when occupants spend time outside of their typical office spaces for meetings, breaks etc. This configuration would also not be appropriate in settings where occupants spend a more equal amount of time in each zone, such as in a domestic application. The house and room total values also showed a reduction on average error rate, although it should be noted that the even the greatest error reduction was only 5%. This is significantly less impact than the Wi-Fi data, likely due to the technical limitations with collecting any Bluetooth data at all from several of the participants.

Table 6-19 - Average RMSE per Zone for ANNs trained with different Bluetooth Beacon Data Inputs – Average Results from 100 Trainings with Phase 2 Data

	Baseline full sets - no BT data	House Total Only	Presence by Occupant	Room Total	Presence assigned to main zone	Presence to main zone, others get all
MGH A01 (meeting space)	0.79	0.86	0.92	0.80	0.80	0.90
'MGH A02' (multi office)	0.84	0.80	0.96	0.80	0.62	0.63
'MGH A03' (kitchen)	0.18	0.19	0.30	0.18	0.18	0.26
'MGH A04' (single office)	0.44	0.44	0.79	0.42	0.41	0.41
'MGH A05' (single office)	0.20	0.20	0.59	0.20	0.73	0.70
'MGH B01' (single office)	0.38	0.34	0.57	0.35	0.41	0.39
'MGH B02' (multi office)	0.51	0.47	0.63	0.52	0.45	0.45
'MGH LG01' (multi office)	1.94	1.73	1.04	1.84	1.41	1.47
MGH LG03' (display space)	0.00	0.00	0.00	0.00	0.00	0.00
Other	0.30	0.30	0.38	0.30	0.30	0.39
Mean	**0.66**	**0.63**	**0.73**	**0.64**	**0.63**	**0.65**

One of the goals of running the Bluetooth Beacon data collection was to assess whether a system such as this could be run in place of a more complex multi-sensor model. The findings suggest that this particular setup is not appropriate to run as a standalone occupancy detection system due to the issues found with signal strength attenuation by bodies and building materials, and due to the same factors may not be a significant contribution to a system attempting to aggregate the number of occupants within a larger space. This data, however, may have value as a contribution to a system attempting to track the location of individual occupants alongside supplementary data.

6.6 Alternative Model Structures

6.6.1 Two-stage Total-then-Distribution Method

In the analysis of the manually counted occupancy data from test phase 1, it was found that the Wi-Fi data from this period provided a good approximation of the total number of people present in the house at a given time, including visitors unknown to

the house, as shown in Figure 6-17. This figure shows the number of occupants detected at signal strength of -60dbm or stronger, with static devices omitted as described in section 5.4.2.

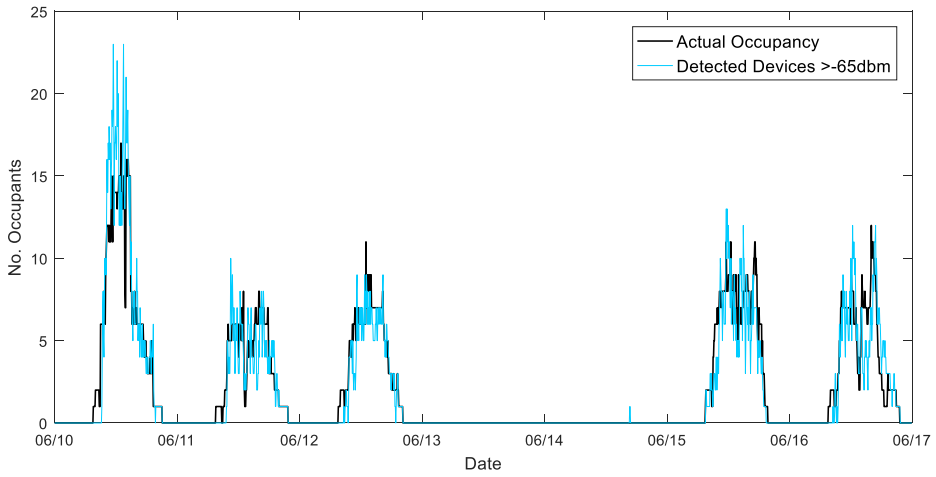


Figure 6-17 - Comparison of Number of People and Number of Wi-Fi Detections, Phase 1 Data

Given the likelihood that some occupants were not carrying a Wi-Fi enabled device and the inconsistent detection of some tested devices, it is unlikely that a model based purely on the Wi-Fi data can attain a perfect estimation of occupancy. However, with the close correlation shown in Figure 6-17, it was proposed that an alternative model structure could give a level of accuracy similar to the all-sensors model, but with a reduced number of sensors needed and amount of computational time needed to train and run an estimation of the number of people per zone.

Figure 6-18 shows the proposed alternative structure: two networks run simultaneously to estimate the total number of people in the building using the Wi-Fi data and the distribution of people in the building based on the local CO_2 concentrations. The output of the total occupancy model is constrained to positive outputs only, while the output of the distribution is constrained so that all outputs total to 1. The estimated number of people can then be calculated by multiplying the distribution outputs by the total occupancy.

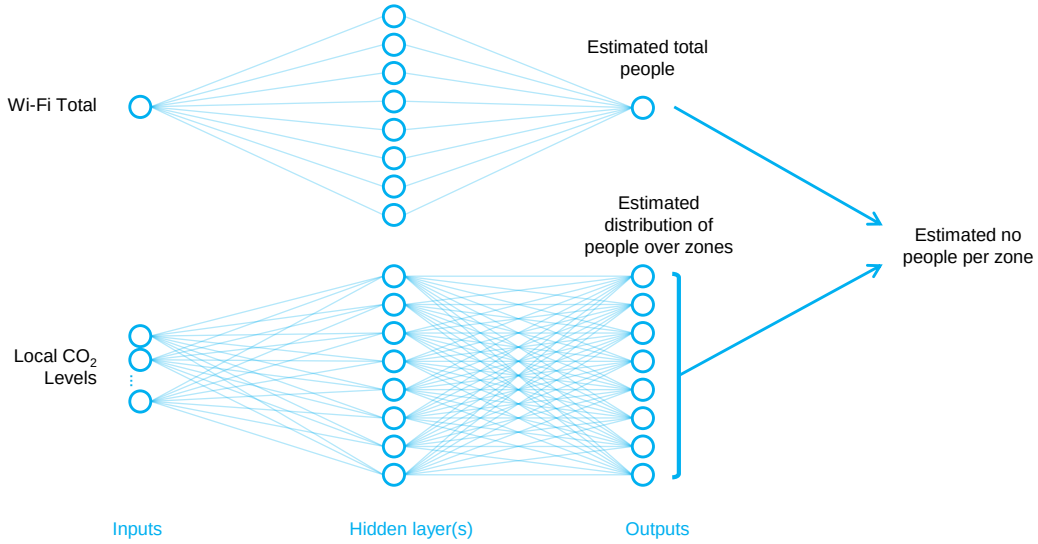


Figure 6-18 – Two-stage Alternative Model Structure to Estimate Occupancy from Wi-Fi and CO_2 levels

The RMSE for this approach against alternative ANN structures is summarised in Table 6-20. For the phase 1 test week, this approach gave an average error that was comparable to the other ANN structures tested, although higher than the reduced-feature set model. It should be noted that Wi-Fi-Distribution method was effective at eliminating false negatives when the building was not occupied, thanks to the heavy reliance on the presence-based Wi-Fi data. It is also worth noting that this method uses a considerably smaller set of sensors and training time, to produce comparable results.

When applied to the Phase 2 data, Table 6-21, this method saw a similar rate of success. The close correlation between the number of detected devices and the number of occupants that was observed in Phase 1 was not seen so consistently in Phase 2, suggesting that the closeness of the correlation may have had an element of coincidence. While in most zones this method produced an increased error relative to the manual feature selection model, it should be noted that the intermittently occupied meeting space A01 saw a lower average error with this simpler model. This suggests that a Wi-Fi-based approach may have value in spaces that cannot be represented well by the physical sensors: those that are occupied intermittently, awkwardly shaped with PIR blind spots or that feature sudden large changes in population. For the building in general, this model type was not experimented with further, as the feature select ANN performed better.

Table 6-20 - Average RMSE of the Wi-Fi Total-Distribution ANN Structure against Alternatives – Phase 1 Data

	Whole-house single model	Whole-House, with Wi-Fi	Individual room models, full set	PCA, 4(/max) features + Wi-Fi	Feature select- 5 features	Total-then-distribution, Wi-Fi and CO2
Research office	1.02	1.14	1.12	0.96	0.92	1.39
Sunspace	1.97	1.32	1.17	1.69	0.43	0.80
PhD office	1.31	1.44	1.26	1.23	0.76	1.30
Single-occ office	0.50	0.49	0.45	0.41	0.37	0.76
Kitchen	0.66	0.59	0.47	0.56	0.29	0.34
Basement outer	0.27	0.33	0.15	0.11	0.10	0.15
Basement inner	0.00	0.00	0.00	0.00	0.00	0.07
Other	0.43	0.50	0.30	0.30	0.25	0.45
Mean	**0.77**	**0.73**	**0.61**	**0.66**	**0.39**	**0.66**

Table 6-21 - Average RMSE of the Wi-Fi Total-Distribution ANN Structure against Alternatives – Phase 2 Data

	Whole-house single model	Whole-House, with Wi-Fi	Individual room models, full set	PCA, 5(/max) features + Wi-Fi	Feature select- 4 features	Total-then-distribution, Wi-Fi and CO2
'MGH A01'	1.17	1.40	0.60	0.73	0.88	0.68
'MGH A02'	1.53	2.05	1.01	0.85	1.00	1.32
'MGH A03'	0.46	0.32	0.24	0.16	0.18	0.16
'MGH A04'	1.12	0.95	0.56	0.50	0.46	0.49
'MGH A05'	0.46	0.45	0.21	0.30	0.18	0.40
'MGH B01'	0.83	0.77	0.41	0.33	0.31	0.33
'MGH B02'	0.77	1.06	0.77	0.56	0.64	0.80
'MGH LG01'	0.72	0.74	1.22	0.66	0.43	0.50
'MGH LG03'	0.00	0.00	0.00	0.00	0.00	0.07
Other	0.48	0.45	0.32	0.31	0.30	0.31
Mean	**0.88**	**0.97**	**0.63**	**0.51**	**0.51**	**0.59**

6.7 Interchangeability of Trained Relationships

One of the factors of the trained model that has been identified in the above sections is how application-specific the optimum solutions of a passive occupancy detection model can be. This can limit the easy deployment of wider-scale systems for occupancy detection as it suggests the need for labelled training data from all zones to be monitored. It was investigated whether the need for training data could be reduced by applying a model trained in one zone to the sensor data from another zone. This was investigated by grouping the zones tested into similar types, which could feasibly be interchanged.

Table 6-22 shows a summary of the zones tested in the second phase of data collection in the Mark Group House. Some zones showed very similar sets of effective sensors, while others of the same type required highly different sensor combinations to most effectively gauge their presence. As a general trend, PIR data and CO_2 trend were valuable in most zones. The absolute CO_2 level was found to be more valuable in multi-occupant spaces, less so in single-occupant spaces.

When compared with the observed occupancy patterns and physical characteristics of the zones, some of the differences in effective sensors can be explained. The single-occupant offices A04 and A05, with similar volumes and usage patterns weighted the local CO_2 trend heavily, while the larger-volume single-occupant office B01 that was occupied more sparsely did not. This could be explained by the shorter average presence duration and the larger room volume, which would mean that the local CO_2 level did not have the time to accumulate a noticeable increasing trend in CO_2 before the occupant had left once again. A similar effect can be seen in the kitchen meeting space A03, which was occupied for an average of 10 minutes at a time and so did not receive a large gain of information from the local CO_2 level.

Table 6-22 – Phase 2 Test Week Occupancy Characteristics grouped by Space Type

	Zone	Max Occ.	Zone Volume (approx.)	Mean Duration	Most Effective Sensors	No. Effective Sensors
Single-Occupant Office	A04	5	27 m^3	94 min	PIR, CO2 Trend	2
	A05	1	27 m^3	106 min	PIR, CO2 Trend	2
	B01	5	68 m^3	46 min	PIR, Ext Door, Windows	4
Multi-Occupant Office	A02	5	63 m^3	217 min	CO2 Trend & Level, PIR, Windows	5
	B02	3	65 m^3	170 min	PIR, CO2 Level & Trend, Windows, RH, Temp	8
	LG01	3	55 m^3	176 min	Ext Door, Wi-Fi Total, PIR	2
Meeting/ Communal Spaces	A01	7	128 m^3	45 min	PIR, CO2 Trend, Windows	9
	A03	2	34 m^3	10 min	PIR, Windows, Ext Door	4
	Other	2	n/a	27 min	Wi-Fi total	1

Given the dependency on physical properties and behavioural usage patterns, it was proposed that spaces with similar values for both of these may be possible to interchange without a decrease in the model performance. The two similar single-occupant office spaces were identified as a possible target for a single-training model that could apply equally to either office. A network was trained on the local PIR and CO_2 trend data from zone A04. This pre-trained network was then supplied with new data points from the corresponding sensors in zone A05, making estimates of the occupancy rate in this zone. The error on this estimate was calculated from the known actual occupancy rate of A05. For comparison, this process was repeated for each of the single-occupancy offices, with the results summarised in Figure 6-19. Here it can be seen that zones A04 and A05 gave the best error rate when trained on their own data, but gave a similarly good error rate when their network was trained on the opposite zone's data. This suggests that rooms with similar physical characteristics and usage patterns could have an effective trained occupancy detection model applied with a significantly reduced need for labelled training data: in offices with large numbers of similar rooms, this relationship would be particularly useful.

In comparison with the similarly sized zones, the larger volume, more sparsely occupied single-user office B01 did not offer an effective training for the other zones, seeing a significantly increased error rate. As a point of interest, the networks trained on A04 and A05 did perform reasonably well in zone B01, although not as well as a

network trained on the larger set of optimum sensors found for zone B01 during feature selection.

It should be noted that the tests in this section were conducted on the winter-climate Phase 2 data only, as some of the single-occupant offices were omitted from Phase 1 due to lack of manual occupant location data. This means that highly user-specific behavioural data such as the window opening patterns during summer was not factored into this test. It can be expected that a trained model that relies on more heavily behavioural data streams would have less success in application to other spaces populated by different users. This effect is seen somewhat with the decreased performance in the more sparsely occupied zone B01, and further highlights the application-specific nature of the trained multi-sensor model approach.

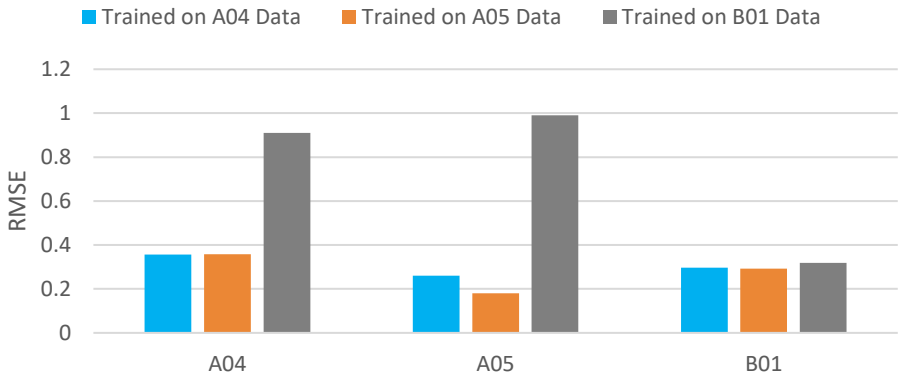


Figure 6-19 - Interchangeability of Networks Trained on Single-Occupant Office Data

Figure 6-20 and Figure 6-21 show the results of a similar process applied to the multi-occupant offices and meeting spaces, respectively. Due to the varying set of ideal sensors for each of these zones, these tests were conducted on sets of sensors that showed a high information gain on most of the zones in the category. This meant that none of the zones were trained on the exact optimum sensor set as found in the feature selection stages. It was expected that the networks trained on other zones' data would be less effective than those trained on data from the same zone, due to the larger differences in the physical characteristics, maximum occupancy and use patterns of these spaces. This effect was seen across most zones, reinforcing the finding that passive occupancy detection can be highly dependent on the properties of the space being monitored.

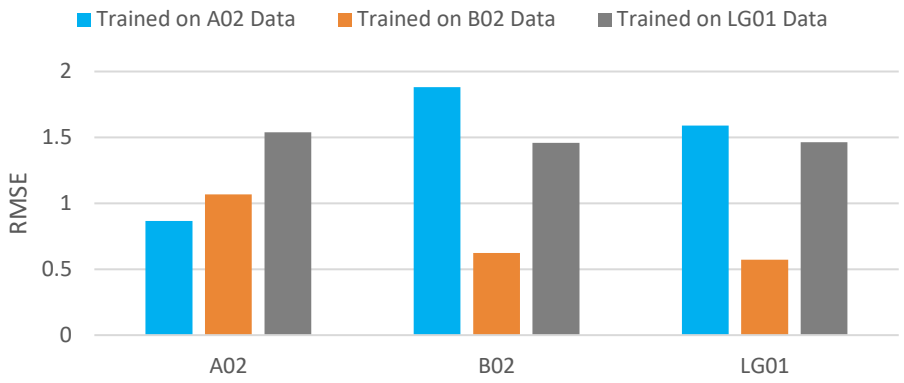


Figure 6-20 - Interchangeability of Networks Trained on Multi-Occupant Office Data

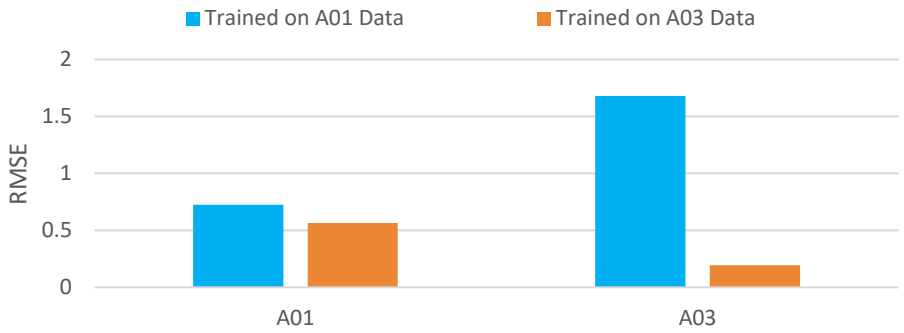


Figure 6-21 - Interchangeability of Networks Trained on Single-Occupant Office Data

It should be noted that the tests on the interchangeability of trained networks were limited to a small sample size of each space type in this application. A more comprehensive picture of the relationship between usage patterns, physical characteristics and the relationship to measured physical properties could be made in a study with a greater number of examples of rooms with similar properties.

6.8 Proposed Model Structure

A final proposed model structure was developed using the combined findings from the sections above. Given the marked improvement on the model accuracy made by including the PIR data, the final proposed model includes PIR in each zone as a requirement. As this was not available in the Phase 1 test data, the following is based on the Phase 2 data onwards. The models were constructed with the following criteria:

- Room-level ANNs, with a separate network trained for each monitored zone.
- Feature Selection to omit unnecessary inputs, driven by the information gain estimated by training networks on individual sensors.
- Pre-processing of CO_2 level to include noise-reduced trend data as an input

- Wi-Fi total number of devices included as an input, selected individual users not included due to the erroneous assignment seen in testing.
- Levenberg-Marquardt training, as this was found to be the most consistent in testing.

The proposed structures for each zone were selected through a directed, cyclical trial-and-error process. The optimum sensors for each zone were ordered according to their individual performance as detailed in the previous sections, with the number of features included in each zone determined by the lowest error rate observed. The optimum hidden layer structure was selected by testing a range around the initial 10-neuron baseline structure and testing further structures similar to the most successful result. Through several iterations of selecting the most successful combination of structure and inputs, the proposed detection model per zone is summarised in Table 6-23.

Table 6-23 - Summary of Proposed Detection Model per Zone

	Optimum Hidden Layer Structure	Optimum No Features	Features	Overall RMSE on test days
'MGH A01'	[5 2 2]	2	PIR, CO_2 Trend	0.42
'MGH A02'	[10]	4	CO_2 Trend, CO_2 Level, PIR, Window 1	0.66
'MGH A03'	[5 2 2]	3	PIR, Window 1, Window 2	0.16
'MGH A04'	[5 2]	2	PIR, CO_2 Trend	0.34
'MGH A05'	[5 2]	2	PIR, CO_2 Trend	0.16
'MGH B01'	[5 2]	4	PIR, Ext Door1, Window 1, Ext Door 2	0.28
'MGH B02'	[5 2 2]	5	PIR, Wifi Total, CO_2, CO_2 Trend, Humidity	0.50
'MGH LG01'	[5 2]	3	Wifi Total, Ext Door, CO_2 Trend	0.45
Other	[5 2 2]	1	Wifi Total	0.30

6.8.1 Multiple Models to Reduce Overfitting

While the most successful sensor sets and model complexities were found, some zones still showed some indications of overfitting when presented with new data points. It was investigated whether a model combination approach, as described in section 6.5, could further improve the error on the test data. By this method, a set of networks trained on the same data that have found slightly different local minima

should produce a lower error if their outputs are averaged, as the individual failings of each network should be cancelled out to some extent by the other networks.

For each zone, twenty separate neural networks were trained with the optimum structure as described in Table 6-23. Each of these networks produced a slightly different set of outputs, due to the tendency to fall into local minima. Table 6-24 shows the RMSE when these twenty networks were used individually (where the output of each network was used to calculate its own RMSE value) and as a combined model (where the final estimation of the occupancy rate was taken as the averaged output of all twenty networks). It can be seen that the RMSE on the test data was reduced when the models were combined. The final proposed model therefore uses the combination on 20 trained networks for its final output.

Table 6-24 - Comparison of the RMSE of 20 networks when used individually or combined as a group

	A01	**A02**	**A03**	**A04**	**A05**	**B01**	**B02**	**LG01**	**Other**
Avg Individual RMSE of 20 Trainings	0.53	0.70	0.16	0.37	0.17	0.28	0.55	0.48	0.30
RMSE of 20 Combined Trainings	0.42	0.66	0.16	0.34	0.16	0.28	0.50	0.45	0.30

The output from these models is visualised in Figure 6-22 and Figure 6-23. It can be seen that the ability of the model to appropriately represent changes in occupancy varies by zone, with a relatively poor performance seen in zones A03, B01 and the circulation spaces. In each of these zones, the occupancy patterns are characterised by short stays by small groups of occupants, meaning that the measures with a delay such as the CO_2-based inputs fail to respond enough during occupancy to register reliably with the model. There were also known issues with the manual location reporting of B01's main occupant, who gave feedback that the timing of their shorter stays may not have been accurately reported. However, in zones with greater number of occupants and longer periods of presence, the performance of the detection model is shown to be relatively successful. Across all zones during the test days of the training data, the models averaged a RMSE of 0.36, meaning that the average estimation of the local number of people was within less than 0.5 away from the actual number of occupants. It should be noted that individual errors over the testing period did exceed

this value. A sample of the Matlab code used to construct the final models is included in the physical copy of this work in Appendix 10.7.

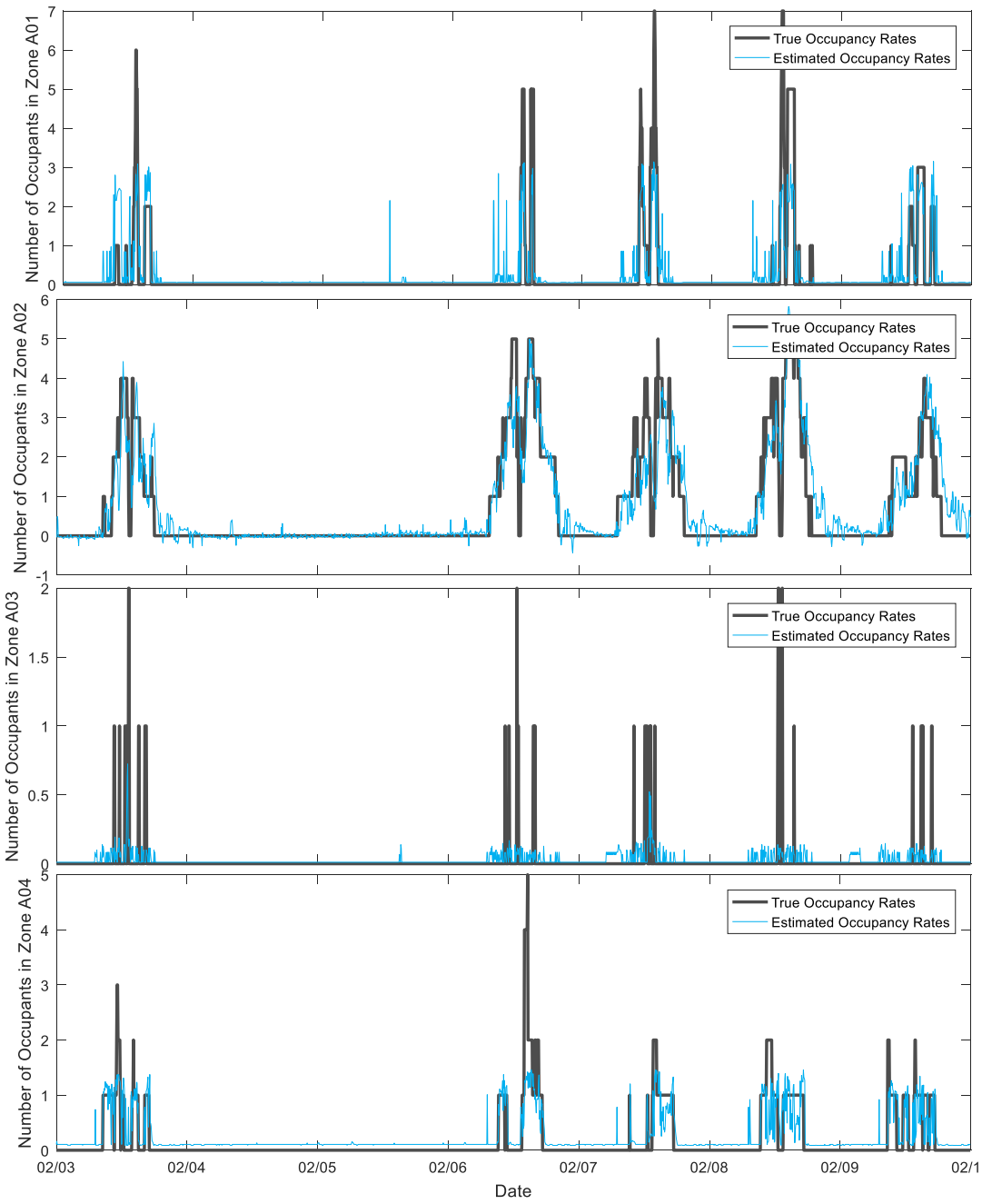


Figure 6-22 - Training Week Performance of Proposed Occupancy Detection Models 1-4

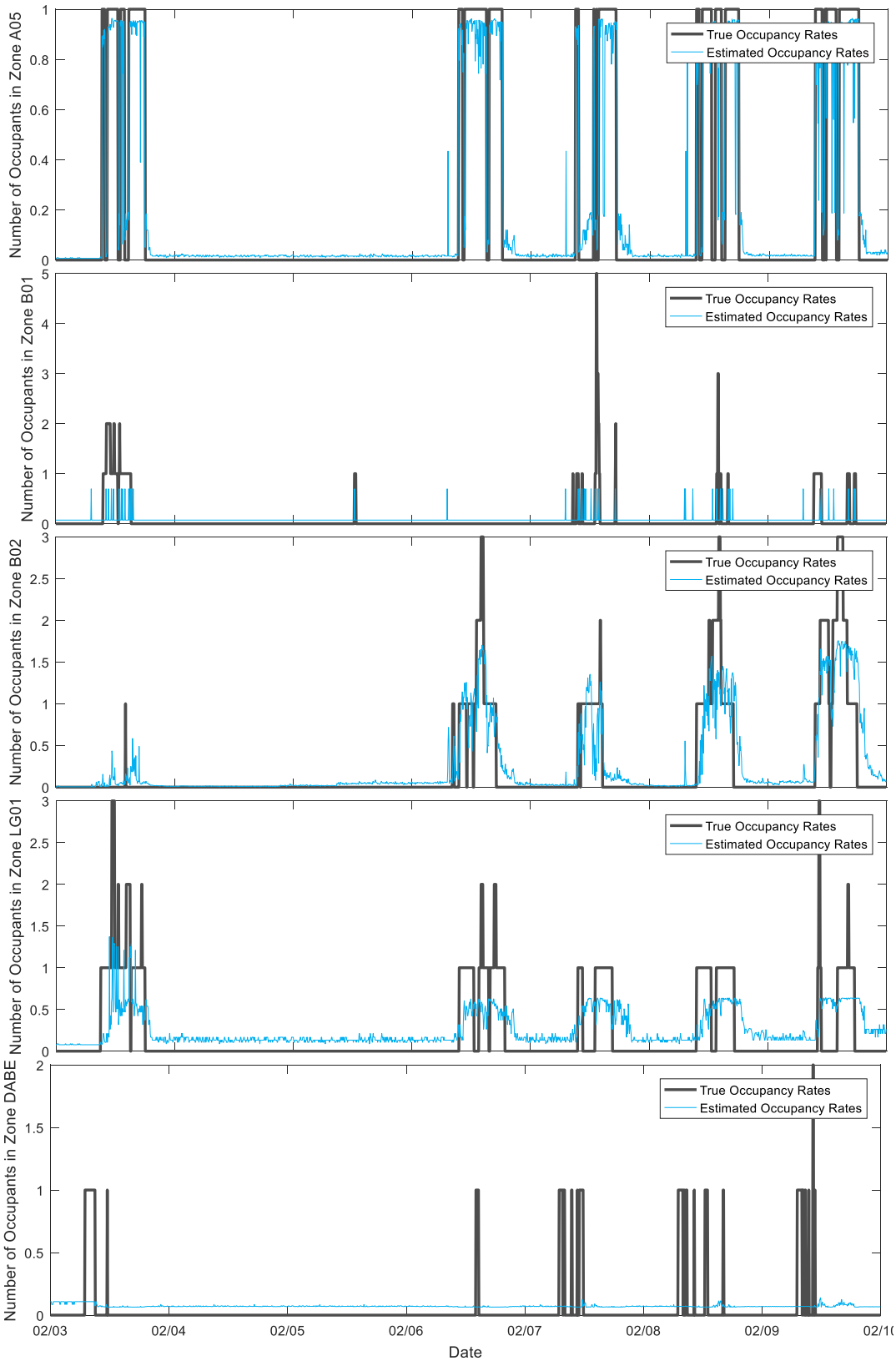


Figure 6-23 - Training Week Performance of Proposed Occupancy Detection Models 5-9

6.9 Conclusions

The focus of this chapter was the development of a model to interpret useable occupancy data from raw sensor data. For this task, machine learning techniques were applied to train a model based on a set of labelled training data. A range of methods were reviewed for suitability, with an Artificial Neural Network selected due to its ability to represent highly nonlinear systems without the need to specify exactly what form the model should take.

Initial testing of the model proved that it is possible to gain more information on the number of people in a zone from a neural network trained on multiple sensors than any one sensor alone can provide. However, including all available data to the model decreased performance relative to a smaller set, meaning that a balance must be found in the exact inputs supplied to the model. A range of techniques to allow for a reduction in the number of inputs to the model were tested. Typical principal component analysis methods were found to be unsuitable for this purpose, as variation in the sensor data was affected by external sources other than occupancy, for example the internal temperature was affected by local weather conditions. Manual feature selection based on a systematic assessment of information gain was found to be a more successful method.

Methods for pre-processing incoming data were tested, with some successes. It was found that including the hourly trend in local CO_2 data provided more information than the absolute CO_2 level alone, particularly in spaces that were not occupied consistently throughout the day. Noise removal on data inputs was found to be effective as long as the noise removal technique did not cause a significant time lag on changes in the data values. Data from personal mobile devices also required some pre-processing to ignore non-personal devices and to reduce the total number of inputs supplied to the model.

One of the major observations of the testing covered in this chapter is that the passive-sensor-based model was highly specific to the characteristics of its application: the effectiveness of sensor types and complexity of pattern recognition possible varied significantly between zones in the same building. Where common trends between spaces were observed, these were still somewhat dependent on the specifics of how the sensor data was measured or processed. For example, the common occupancy-

measurement sensors of local PIR count and CO_2 level were found to be common high-ranking features across multiple zones in this work. However, the Phase 1 PIR data from adjoining circulation spaces was ranked as one of the least valuable features, calling into question the value of single-point PIR sensing strategies as was seen in Chapter 4. The absolute CO_2 level was also ranked as a less useful feature in some spaces, where the CO_2 trend data provided more value. Sensors that were consistently ranked as less useful included the local temperature and humidity.

The variation in effectiveness of sensor data types appeared to be linked to both the physical characteristics and behavioural patterns of occupants, meaning that the same system applied in different spaces could produce significantly different accuracies. It was also shown that a model trained on a limited time period could even change in accuracy over time, as demonstrated with the season-dependent window opening behaviours that were highly effective in the Phase 1 summer model, but not observed at all in the Phase 2 winter model. In application, it is essential that these differences are accounted for if a system or model has only been tested in one type of space and for a limited time period, as is typical in many research applications discussed in Chapter 3. For more comprehensive research into the relationship between space use and the effectiveness of passive occupancy sensing methods, much broader datasets of diverse space types would be needed. It should also be noted that some spaces did not see enough use during the training period to properly encode their patterns into the model: this is a consequence of relying on a relatively small amount of training data.

A significant shortcoming found in this approach was the reliance on manually labelled training data. As discussed in previous sections, due to the multi-occupant nature of the office space tested, it was prohibitively difficult to collect 5-minute resolution location data from all occupants over a longer time period, meaning that the amount of training data available was not sufficient to train a more complex model sensitive to more intricate interactions between the physical sensor data collected and the observed occupancy rate. Models had to be kept relatively simplistic to avoid the overfitting issues associated with a smaller training dataset. However, without some form of manually labelled data, it would be impossible to verify the effectiveness of the model. Where non-trained unsupervised modelling approaches are taken, the model must be manually encoded with assumptions about occupancy events, similar

to the assumptions made in Chapter 4 when examining case study data with no verified context for motion sensor data.

Ultimately, it is expected that as technologies progress, an occupancy detection model based on the particular passive data sources tested in this work may become redundant. For example, if a more reliable system to detect known occupants can be made using data from personal devices, this would perform mostly the same function at a considerably lower time and monetary cost than installing the full range of sensor types tested in this work, and may only require a smaller supplement of additional sensors to detect outlying occupants who do not carry devices. In post-occupancy trend analysis or general pattern observation, there is less penalty if some occupants are missed. However, in building services control it is essential that the building can respond to any occupant. In theory, any reliable occupancy detection system could be used as a data supply for a predictive model for future occupancy, which is developed in the next chapter.

7 OCCUPANCY PREDICTION MODEL DEVELOPMENT

7.1 Introduction & Aims

As discussed in Chapter 6, it was decided that a system to use sensor data to estimate the local number of occupants and produce a short-term prediction of future occupancy rates would be produced in this study in order to allow for improved pre-emptive control of building systems. The task of converting sensor data to a prediction of the number of occupants in a space was split into two modules to allow for changes in available technology on the detection side. The first module is a 'detection' model to convert sensor data into an estimation of the number of people currently in a space in real time, discussed in detail in Chapter 6. This chapter covers the development of the second module: a 'prediction' model to take the number of people over a recent time period and predict the number of people in the future. In order to systematically test this model, a method to generate occupancy datasets with varying pattern complexities was also developed as part of this chapter.

The detection model was designed to be trained on a relatively small length of training data, with the assumption that the relationship between sensor data and the local number of people largely relies on physical properties, and so should stay relatively constant. However, when predicting patterns in local occupancy over time, it was considered more likely that, as building occupants, schedules and room uses change over time, a predictive model trained only once would quickly become outdated and encounter higher errors as patterns diverge from those learned by the model. It was therefore a key aim to produce a model that could feasibly be updated continuously during its operation. It was also decided that the model must reach a reasonable error rate within three weeks of training data, as it is unreasonable to expect a non-functional system to run in a building for months before it begins working as intended.

Another aim for the predictive model was to achieve an improved performance in unusual situations relative to simpler prediction heuristics. For example, a static assumed schedule of full occupancy during office working hours may give a reasonable estimation on most regular days, but would consistently fail in its predictions on a day when most regular occupants are on holiday. As was shown in the case studies presented in Chapter 4, these unusual situations are often associated

with a poor match between energy use and demand, leading to a higher than typical energy waste or discomfort of the building users. This makes unusual occupancy situations a primary target for systems aiming to reduce energy waste and so these were considered a priority for the prediction model.

As with the detection model, the prediction model was based on data from the Mark Group House testbed as described in section 5.2. The model was tested for its long-term performance using a generated dataset designed to have the same characteristics as the Mark Group House data.

7.2 Selection of Appropriate Machine Learning Methods

Similarly to the detection model, the prediction model was treated as a regression problem, allowing for prediction outside the scope of the initial training data. However, while the sensor data and number of occupants could be treated as a series of discrete samples, the prediction model was required to deal with time series data, where the continuity from one sample to the next is considered. While the machine learning methods discussed in Chapter 6 are still mostly applicable in this case, some variations are required to properly process time series data.

A neural network approach was preferred for coherency: if needed, the ANNs for detection and prediction could technically be combined and run as a single network after training, in order to streamline the process of running at each 5-minute timestep. The following time-series appropriate methods were considered:

7.2.1 Nonlinear Input-Output

This structure, similar to the detection model, would simply take a set of inputs and be trained towards the required output, as illustrated in Figure 7-1. Data from the current time y_0 is provided as an input. Data from previous timesteps could also be manually formatted and provided as additional inputs to allow the network to account for the recent past. If the network is needed to predict multiple steps into the future, this could be achieved by including multiple outputs, as in the figure.

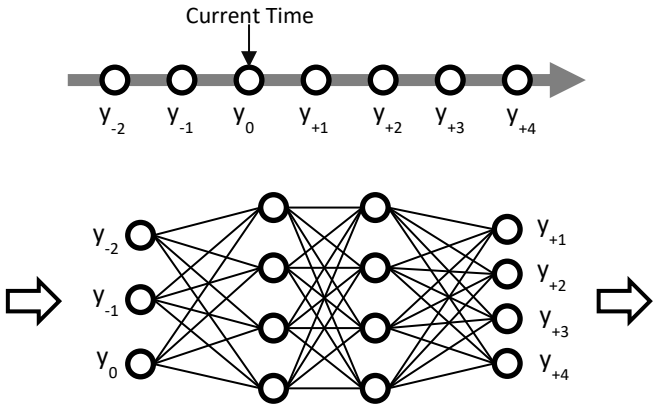


Figure 7-1 - ANN Time Series: Nonlinear Input-Output

A disadvantage with this method is that the data would have to be manually restructured into appropriate chunks at each 5-minute timestep for both the training and operation of the model.

7.2.2 Nonlinear AutoRegressive (NAR)

A recurrent neural network can be used to automate the process of including data from previous timesteps. Recurrent networks such as the Matlab Toolbox Nonlinear AutoRegressive direct the output at each timestep to feed back into the network at the next timestep, as shown in Figure 7-2. This creates a set of 'delay' inputs from previous timesteps, shown in grey in the figure. If the network is needed to predict multiple timesteps into the future, this is typically achieved by iteratively feeding a single-timestep predicted value back into the model multiple times, at each iteration receiving a prediction of one step further into the future. This means that predictions further into the future are based more heavily on assumptions from previous timesteps, meaning that errors could compound quickly.

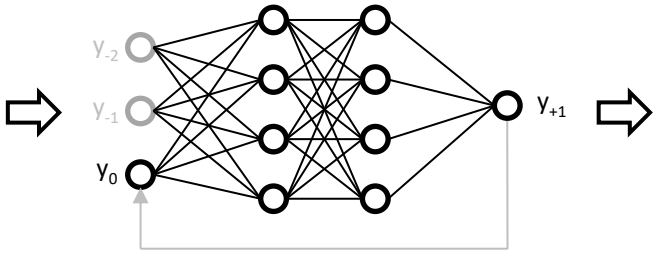


Figure 7-2 - ANN Time Series: Nonlinear AutoRegressive

7.2.3 Nonlinear AutoRegressive with eXternal input (NARX)

Similarly to the NAR structure, the NARX (Nonlinear AutoRegressive with eXternal input) Neural Network structure automates the data handling of time series. The output is returned as feedback to the inputs for the next timestep. This structure also

includes additional input variables that are not part of the output feedback loop – denoted in the figure by x_0, x_{-1} etc. In this case, the external inputs could include contextual information on the timestep in question, such as the time of day or day of week. This allows the network to better represent patterns that rely on factors other than the output variable. In this study, it is expected that the occupancy patterns will rely heavily on time-related factors, and so this structure was chosen for the prediction model.

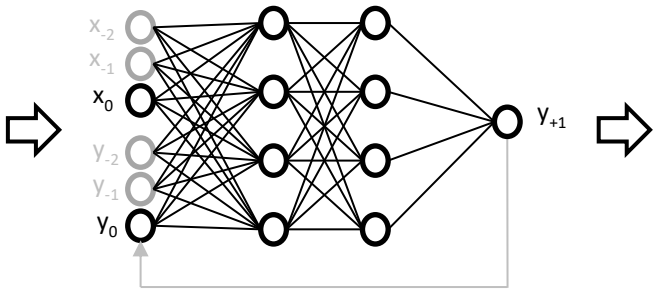


Figure 7-3 - ANN Time Series: Nonlinear AutoRegressive with eXternal input

7.3 Accounting for Uncertainty in Models

With the prediction of future events, it is known that the model will never have a perfect accuracy, as there will always be some unexpected variability to human behaviours. It is therefore of particular interest to be able to provide some measure of confidence in the model output, where unusual situations can be identified and potentially poor predictions will not be treated in the same way as those with higher certainty. The confidence of the model output, or a potential range of expected outputs, can be obtained through a range of methods. This section discusses some of these methods.

7.3.1 Combining models

As was introduced in Chapter 6, the combination of multiple models trained on largely the same dataset can provide several benefits, including the reduction of overfitting issues [215]. This technique can also provide a range of potential outputs from the same input data by considering the variation in each model's output. Where all models in the group provide similar answers, the output might be considered more certain. Where models give highly varied answers, the output is less certain, likely due to a lack of similar situations in the training set or a high variability of outcomes from similar situations in the training set.

7.3.2 Mixture of Experts

This method requires the training of several models as above. However, the spread of models is more structured, with each model intended to specialise in a certain aspect of the data. To define these aspects, the data can be clustered into groups of similar input-output relationship (not just similar inputs or similar outputs). Nets are then weighted according to which aspects they address. With smaller training sets, this technique can have limited benefits, due to a lack of examples to cluster in order to identify the aspects targeted in the model training. As the system in this study is designed to start training on a relatively small initial training set, this method was deemed inappropriate.

7.3.3 Bayesian ANN

Producing similar results to combining models, Bayesian methods seek to account for uncertain data by producing a distribution of potential predictions (whether formally solved or via sampling). This allows for more effective combined prediction compared to a single model, and also allows confidence in the predicted value to be evaluated using the spread of answers from the distribution [218].

For all but the simplest models/networks, Bayesian ANN methods are not feasible to solve fully, as the computational need can be intensive and increases exponentially with the number of hidden units [219]. Instead, sampling methods such as Markov Chain Monte Carlo (MCMC) sampling can be applied.

The premise of Bayesian treatment of Neural Networks is to assess the likelihood of any possible values of the network's parameters when given the training data. The training is started with an assumed 'prior' distribution over the possible parameters, which can be assigned based on intuitive knowledge or standard rules to initiate the process.

The prior is then adjusted to reflect the given training data points, which make certain configurations of parameters more or less likely. The adjusted probability distribution is called the posterior. With enough training data, the posterior is guided towards favouring values of the parameters that explain the training data well.

The posterior distribution can then be used to give an estimation of the confidence in the output in two ways:

Multiple models: A range of models can be generated using values of the parameters that comply with the posterior distribution. When given new input data for testing, these models should produce similar outputs where the model is certain and more varying outputs where the model is less certain, as with the combined models approach. A single-value prediction output can be taken from the outputs using an average weighted by the posterior distribution, or a predicted distribution can be inferred from all the model outputs.

Gaussian Distribution: The confidence in the model output can also be estimated as a Gaussian distribution, with a mean of the actual model output and a standard deviation that represents the outer confidence limits of the model output. Here, a larger standard deviation on the output for some particular input shows that the model is less certain of this output, and a smaller standard deviation indicates a higher level of certainty. This method does include an assumption that the range of outputs from the model would take a Gaussian form, which may not always be appropriate. However, this method avoids the need to calculate the output from large range of models.

7.3.4 Selected Method – Bayesian Neural Network

Although the methods requiring the training of multiple models have seen high levels of success in other studies, they were considered less appropriate for this study's application. As the prediction model is intended to be re-trained continuously during its operation, any method that requires the training of a large range of models becomes less feasible in terms of computational time, especially as the training dataset becomes larger and larger over time. It was decided that the Bayesian ANN method would be tested, using the confidence levels computed by assumption of a Gaussian distribution, once again to minimise computational stress on a system intended to operate in real time.

In direct comparisons of machine learning methods, Bayesian Neural Networks using the techniques found in the Matlab NN toolbox were one of the most effective solutions alongside Gaussian Processes [220] or compared favourably to most other methods [213]. It was assumed that this method would produce a reasonable fitting performance as well as providing the uncertainty estimation.

The remainder of this section provides a more detailed explanation of the BNN method and a derivation of the method used to calculate the confidence levels.

Where most machine learning methods seek to minimise the error over the training set, Bayesian Neural Network treats the problem in a probabilistic manner. That is, all possible weight and bias values for the network are treated as a distribution of n-dimensions, where n is the total number of weights and biases in the network. Given a set of training data, it can be assumed that some values of the weights and biases will be more likely than others. This introduces the concept of the probability of weights, given the training data as a probability distribution across all possible weights.

Network training then becomes a problem of finding the maximised probability of weights given the data. As mentioned above, there are two potential approaches to this problem:

- Train a set of models from within the distribution via application of numerical sampling techniques such as MCMC. The overall model output can then be taken as the average of the trained sub-models, and an uncertainty measure can be quantified by taking models at the 5^{th} and 95^{th} percentile. This technique is not supported in native Matlab applications, although some 3^{rd} party software exists.
- Treat the optimisation in a similar manner to error minimisation, using similar learning algorithms (e.g. Levenberg-Marquardt optimisation) to find the optimal weight distribution. It is assumed that the solution found will be global, although in some applications it should be noted that local minima are found.

Bayesian functionality in the Matlab NN toolbox utilises the second approach, based on the work of MacKay [221], Foresee and Hagan [222]. The derivation described in the following sections is based upon these works. With the Matlab toolbox functionality alone, the Bayesian method is used as a means of network regularisation as follows:

The core principle of the Bayesian method is Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad (1)$$

Here, the posterior distribution $P(A|B)$ describes the probability distribution of random variable A, given information B. This posterior distribution is informed by a prior belief on the probability of $P(A)$, the likelihood $P(B|A)$ and the probability of B, $P(B)$. In essence, the posterior distribution can be considered an updated version of the prior, when given further information.

The aim of Neural Network training is to minimise the objective function *F*, typically a measurement of the magnitude of errors on the training set such as the mean squared error, E_D. Weights/biases of the system are adjusted to minimise this error. Regularisation also seeks to minimise the complexity of the NN model, which can be measured by the sum square of all weights/biases in the model E_W. A regularised objective function can thus be represented by:

$$F = \beta E_D + \alpha E_W \quad (2)$$

A Neural Network model can be described in probabilistic terms if the weights/biases of the model are treated as a random variable $\boldsymbol{W}$, which can take values $\boldsymbol{w}$. Training the model involves finding a set of weights/biases that allow the model to best describe the observed data. Using Bayes' rule, this can be represented as:

$$P(\boldsymbol{w}|D, H_i) = \frac{P(D|\boldsymbol{w}, H_i)P(\boldsymbol{w}|H_i)}{P(D|H_i)} \quad (3)$$

Where D represents the network training data and H_i represents the particular NN model hypothesised. Figure 7-4 shows a visualisation of the relationship between the prior (dotted line) and posterior (solid line). It can be seen that the ideal update from prior to the posterior decreases the width of the probability distribution, representing a decrease in uncertainty on the location of the optimum set of weights $\boldsymbol{w}_{MP}$.

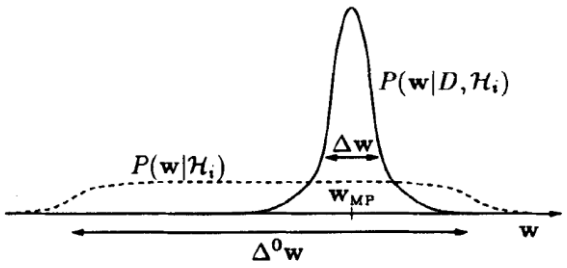


Figure 7-4 - Visualisation of the relationship between Prior and Posterior Distributions [221]

The hypothesised model H_i can be broken down into components M: the model structure/basis functions chosen, and α and β: the regularisation parameters. Bayes rule is thus represented as:

$$P(\boldsymbol{w}|D,\alpha,\beta,M) = \frac{P(D|\boldsymbol{w},\beta,M)P(\boldsymbol{w}|\alpha,M)}{P(D|\alpha,\beta,M)} \quad (4)$$

Here, there is an assumption that noise on the training set data is Gaussian and prior distribution for weights is Gaussian. Using the probability density function for Gaussian distributions allows the likelihood function and prior to be calculated as:

$$P(D|\boldsymbol{w},\beta,M) = \frac{1}{Z_D(\beta)}\exp(-\beta E_D) \quad (5)$$

$$P(\boldsymbol{w}|\alpha,M) = \frac{1}{Z_W(\alpha)}\exp(-\alpha E_W) \quad (6)$$

Where $Z_D = (\pi/\beta)^{n/2}$, $\beta = 1/\sigma_v^2$, $E_D = \sum_m \frac{1}{2}[y(x_m) - t_m]^2$, $Z_W = (\pi/\alpha)^{N/2}$, α is the regularising constant and E_W is the sum squared of the network weights/biases. Other potential choices exist for the prior, as described in [221], but this project will be focusing on the equations as above, as used in the Matlab toolbox.

The posterior can now be represented by:

$$P(\boldsymbol{w}|D,\alpha,\beta,M) = \frac{\frac{1}{Z_W(\alpha)}\frac{1}{Z_D(\beta)}\exp(-\beta E_D - \alpha E_W)}{Normalisation\ Factor} = \frac{1}{Z_F(\alpha,\beta)}\exp(-F(\boldsymbol{w})) \quad (7)$$

Where $Z_F(\alpha,\beta) = \int d^k\boldsymbol{w}\ \exp(-M)$. In application, this is typically difficult or impossible to calculate directly, but can be estimated via Taylor series expansion [222].

The weights at the maximal probability $\boldsymbol{w}_{MP}$ can be found using optimisation methods as discussed in previous sections, as minimising the objective function should be equivalent to maximising the probability.

With known values of $\boldsymbol{w}_{MP}$, α and β, the model can then be used to make predictions from new data inputs. This is as far as the default Bayesian method is applied in the Matlab ANN toolbox at the time of writing.

The confidence in the BNN predicting a target value t when given a new set of inputs $\boldsymbol{x}$ and weights $\boldsymbol{w}$ can be modelled as a Gaussian distribution [223] [224]. Here, the most probable output (Gaussian mean) is the model output $y(x, w_{MP})$ using the maximum probability weights. The variance is given by β^{-1}, which is related to the level of noise assumed on the training data.

$$p(t|\boldsymbol{x}, \boldsymbol{w}, \beta) = N(t|y(\boldsymbol{x}, \boldsymbol{w}), \beta^{-1}) \quad (8)$$

Marginalising with respect to $\boldsymbol{w}$ gives the predicted output for inputs $\boldsymbol{x}$ in terms of the weight-dependent model output and the posterior distribution:

$$p(t|\boldsymbol{x}, D) = \int p\big(t\big|y(\boldsymbol{x}, \boldsymbol{w})\big)p(\boldsymbol{w}|D)d\boldsymbol{w} \quad (9)$$

This evaluation would involve calculation of the model output for every possible set of weights. Once again, this is not a computationally viable evaluation for a neural network with a nontrivial number of weights. To simplify, two assumptions are applied:

- The objective function is approximated locally to $\boldsymbol{w}_{MP}$
- The gradient between network outputs for weights close to $\boldsymbol{w}_{MP}$ is approximated as linear, allowing network output at $\boldsymbol{w}$ to be evaluated quickly

For the first assumption, the objective function for weights local to $\boldsymbol{w}_{MP}$ is described by:

$$F(\boldsymbol{w}) \approx F(\boldsymbol{w}_{MP}) + \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_{MP})^T \boldsymbol{A}(\boldsymbol{w} - \boldsymbol{w}_{MP}) \quad (10)$$

Where $\boldsymbol{A} = \beta\nabla\nabla E_D + \alpha I$. $\nabla\nabla E_D$ can also be called the Hessian of the error function, and represents a matrix of second derivatives of the error function with respect to the weights.

For the second assumption:

$$y(\boldsymbol{x}, \boldsymbol{w}_{MP} + \Delta\boldsymbol{w}) \approx y(\boldsymbol{x}, \boldsymbol{w}_{MP}) + g^T \Delta\boldsymbol{w} \quad (11)$$

Where $g = \nabla y(\boldsymbol{x}, \boldsymbol{w})|_{\boldsymbol{w}=\boldsymbol{w}_{MP}}$, a vector of partial gradient of the network output with respect to the weights, taken at the weight set $\boldsymbol{w}_{MP}$. A single element of this gradient vector is plotted for illustrative purposes in Figure 7-5.

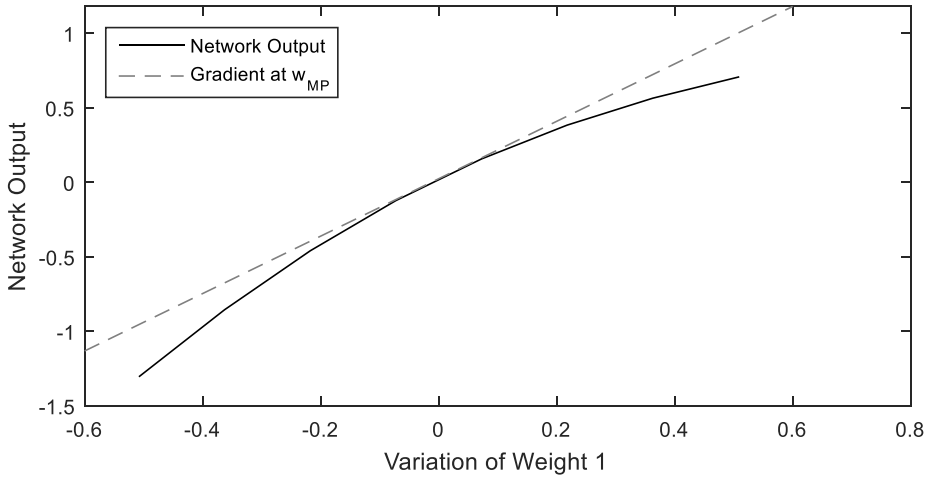


Figure 7-5 - Estimated Network Output Gradient, varying First System Weight Value (data sampled from network trained later in this chapter)

Using assumptions above, equation 8 can be expressed as:

$$p(t|\boldsymbol{x}, \boldsymbol{w}, \beta) \approx N(t|y(\boldsymbol{x}, \boldsymbol{w}_{MP}) + g^T(\boldsymbol{w} - \boldsymbol{w}_{MP}), \beta^{-1}) \quad (12)$$

This expression can then be marginalised with respect to $\boldsymbol{w}$, leaving:

$$p(t|\boldsymbol{x}, D, \alpha, \beta) = N(t|y(\boldsymbol{x}, \boldsymbol{w}_{MP}), \sigma_t^2(x))$$

or

$$p(t|\boldsymbol{x}, D, \alpha, \beta) = \frac{1}{\sqrt{2\pi}\sigma_t} \exp\left(-\frac{\left(t - y(\boldsymbol{x}, \boldsymbol{w}_{MP})\right)^2}{2\sigma_t^2}\right) \quad (13)$$

Where $\sigma_t^2(x) = \beta^{-1} + g^T \boldsymbol{A}^{-1} g$. The variance of the Gaussian σ_t^2 represents the level of uncertainty in the model prediction – with a higher variance showing that the model is less certain about this output. This means that the uncertainty term depends on both the noise on the training dataset (β^{-1}) and the variability of network output under changing weights ($g^T \boldsymbol{A}^{-1} g$). The second term is dependent on $\boldsymbol{x}$ value.

7.4 MATLAB Uncertainty Implementation

It can be seen above that, in order to calculate the variance, and so the confidence level, the following quantities must be known:

- $\boldsymbol{x}$: matrix of network inputs for which the error bars are to be calculated
- α, β : parameters of regularisation, set during the network training process
- $\boldsymbol{A} = \beta \nabla\nabla E_D + \alpha I$: regularised Hessian matrix of network errors on the training set
- g – the vector of partial gradient of the network output with respect to the model's weights

The Bayesian Regularisation training process featured in the MATLAB ANN toolbox (version supplied with Matlab 2015a) is governed by the function "*trainbr*", following the process described above and in the work of Foresee [222]. During this process, the Hessian of the network error $\nabla\nabla E_D$ is approximated internally, using the squared Jacobian matrix (first order derivative). This is considered a valid approximation for the purposes of this work. The parameters α and β are also evaluated as part of the training process.

Calculation of the error bars was therefore a matter of extracting these values from the "*trainbr*" function during training. This was achieved by editing the local copies of relevant MATLAB functions to output the relevant data alongside their normal outputs.

The gradient g at the final trained weight values was numerically calculated by:

- Unpacking trained network weights.
- Making a matrix of varied network weights, including a small increase and decrease on either side of the final weight values $\boldsymbol{w}_{MP}$.
- Repacking each set of varied weights into the network and calculating network output for these varied weights.
- Calculating the linear gradient of network output from two points either side of $\boldsymbol{w}_{MP}$.

7.5 Dataset Generation for Training over a Long Time Period

In order to fully test the long-term performance of a continually updating model, as well as to test the model's response to different complexities of patterns in occupancy data, a dataset was needed that could produce data following the patterns of occupancy observed in the test space, over a time period of arbitrary length.

The data characteristics required were:

- Single-occupancy and multi-occupancy spaces within the same building.
- Known recurring staff with an assigned main zone of the building.
- Unknown visitors to the building.
- Varying occupancy profiles per person, grouped into sets with common features (e.g. teaching staff more likely to have regular weekly absences for lecturing, start time of PhD students more likely to be variable).

Within this structure, the following features were identified in order to generate a realistic dataset:

- Daily patterns – actions dependent on a certain time of day (occupancy profiles)
- Weekly patterns - actions dependent on the day of the week (weekday/weekend, adjusted leaving times on a certain day)
- Date-based patterns - actions dependent on the day of the year
- Random low-level variation – variation on the start/end times of occupancy about the mean value determined by occupancy profiles
- Long-term absences – holidays (some regular, some variable)
- Short-term absences – day meetings, sub-day meetings
- Long-term variations – changes of staffing or of zone associated with an agent over the long term.

A set of occupant or 'agent' profiles were constructed designed to mimic the behaviours of different groups observed in the test building. Table 7-1 provides a summary of how each of the pattern types was implemented, with Table 7-2 providing the specific values used to draw different agent characteristics within each profile group. It should be noted that for all of the profiles, overlapping time periods of absence were aggregated and long-term absences were biased towards starting on

Mondays and/or ending on Fridays. The plain-text algorithm used to generate this data is shown in Appendix 10.6, with the full Matlab code used included in the physical copy of this work in Appendix 10.7.

Table 7-1 - Generated Occupancy Profile Characteristics

Pattern Type	**Method**
Main Zone	Zone assigned to each agent dependent on occupant type.
Regular Daily Profile	Standard profile 9:00-18:00. Variation +/- in hours drawn from Gaussian for each agent
Regular Lunch Profile	Standard profile 13:00-14:00.
Daily Variations	Regular daily events that interrupt the standard profile. Number of events drawn from exponential distribution. Start and end times randomly drawn within the 9-18 range for most occupants.
Weekly Variations	Regular weekly events that interrupt the standard profile. Drawn as above. Day of the week randomly assigned to each event.
Monthly Variations	Regular monthly events that interrupt the standard profile. Drawn as above. Day of the month randomly assigned to each event.
Date-based patterns/Long-term absences	Common holiday periods drawn in three parts of the year for each agent. Spring holidays of all agents confined to include a common 'Easter' weekend. Summer holidays drawn of Gaussian length, with Gaussian start date and adjustment to bias towards start/end on Mon/Fri. Winter holiday constrained to include 25 Dec – 01 Jan.
Random low-level variation	Standard deviation values for random variation for each agent drawn from Gaussian distribution.
Random weekly variation	Each agent assigned a likelihood of presence at weekends, randomly drawn within a range specified based on occupant type. A number of unknown visit events each day drawn from Gaussian, number of people drawn from exponential distribution and rounded, zone randomly assigned from subset.
Short-term absences	Likelihood of non-regular day-long absences randomly drawn from range specified based on occupant type. Mean number of non-regular daily short absences drawn from Gaussian for each agent.
Long-term changes	Staff agents reassigned weekly patterns at typical university term times. Random reassignment of a completely new agent determined by given frequency.

Table 7-2 – Values used for Pattern Formation for each Occupant Type

Pattern Type	**Profile 1** (Teaching Staff)	**Profile 2** (Researcher)	**Profile 3** (PhD Student)	**Profile 4** (Cleaning staff)	**Profile 5** (Regular visitor)	**Profile 6** (Unknown Visitor)
Main Zone	Single-occupant zone for each agent	Shared within group subset	Shared within group subset	No	Shared within group subset	No
Regular Daily Profile	mean = 0 start time std dev = 0.5 end time std dev = 0.5	mean = 0 start time std dev = 0.5 end time std dev = 1.5	mean = 0 start time std dev = 1 end time std dev = 2	Regular profile 07:00-08:30.	Random start time between 9-17. Gaussian length mean 1 hr std dev 2	
Daily Variations	Mean = 0.5	Mean = 0.5	Mean = 0.5		Can only enter assigned zone when already occupied	
Weekly Variations	Mean = 6	Mean = 2	Mean = 0.5			
Monthly Variations	Mean = 4	Mean = 4	Mean = 0.5			
Date-based patterns/Long-term absences	Length mean = 10, std dev = 7 Summer date mean = 196, std dev = 35	Length mean = 10, std dev = 7 Summer date mean = 196, std dev = 35	Length mean = 10, std dev = 7 Summer date mean = 196, std dev = 35	Length mean = 10, std dev = 7 Summer date mean = 196, std dev = 35	Length mean = 10, std dev = 7 Summer date mean = 196, std dev = 35	
Random low-level variation	Mean = 0 Start std dev = 0.1 End std dev = 0.5 Lunch std dev = 0.1	Mean = 0 Start std dev = 0.1 End std dev = 0.5 Lunch std dev = 0.1	Mean = 0 Start std dev = 1 End std dev = 1 Lunch std dev = 0.1	Mean = 0 Start std dev = 0.1 End std dev = 0.5	Mean = 0 Start std dev = 1 End std dev = 1	
Random weekly variation	Range 0-0.125	Range 0-0.33	Range 0-0.1			No daily events mean 2 std dev 1.2. No visitors mean 1.5
Short-term absences	Likelihood Range 0-0.25 Short event mean = 1.4 Short event std dev = 0.8	Likelihood Range 0-0.1 Short event mean = 0.8 Short event std dev = 0.8	Likelihood Range 0-0.17 Short event mean = 0.5 Short event std dev = 0.5			
Long-term changes	Frequency of long-term changes: 500 days	Frequency of long-term changes: 500 days	Frequency of long-term changes: 500 days		Frequency of long-term changes: 500 days	

An example of data generated by the process described above is shown in Figure 7-6, where the total occupancy of the building is plotted on the upper graph, with each

individual agent's presence in the building plotted in the lower graphs. It can be seen that this process creates a feasible occupancy profile, with most occupants keeping to a regular, but slightly varying pattern over the long term.

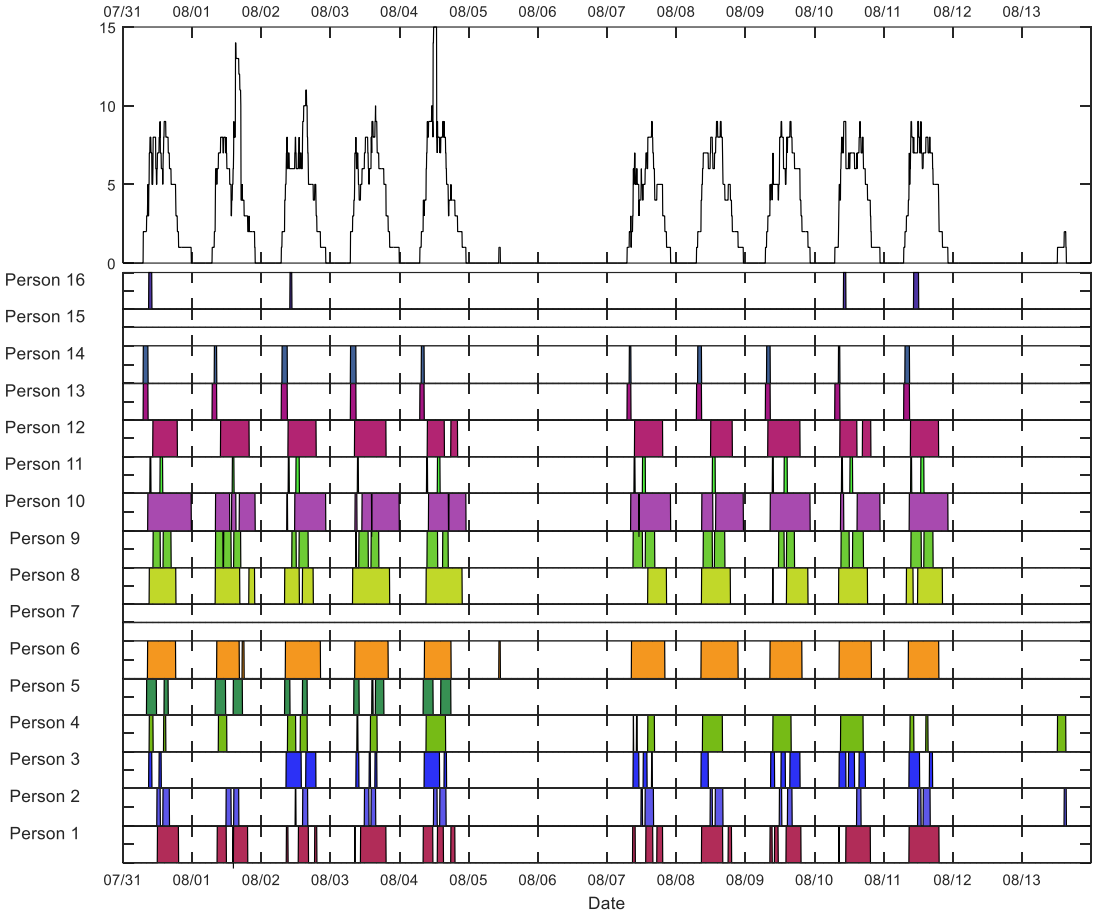


Figure 7-6 - Example of Generated Data for 14-day period

7.6 Testing Model Structure

As the detection model operates on a 5-minute interval, a predictive model that works on the same data needs to predict multiple timesteps into the future to provide useful data, as anticipating the next 5 minutes alone is not especially valuable for the control of building systems with a slow response time. Several methods to predict multiple timesteps into the future were considered for this model. Each of these models was tested on a 50-day occupancy dataset generated with the algorithm described in section 7.5.

7.6.1 NARX

As described above, the NARX network structure is recurrent, meaning that to predict multiple timesteps into the future at time *t*, the output prediction for *t+2* is typically

produced iteratively by producing a single-step prediction for *t+1* and feeding this prediction back into the model. One of the concerns with this approach was that without a defined way to signify to the model that the *t+1* value is an assumed prediction rather than a known observation; poor predictions would be quickly compounded to lead to large errors over longer prediction horizons. The standard NARX structure was tested for comparison with two alternatives that could reduce this compounding error issue.

Figure 7-7 shows the absolute error averaged across a 12-timestep prediction horizon over a 50-day training on the generated dataset, with the network re-training to account for new data at the end of each day. It can be seen that, while the error on the predicted number of people was generally low, the system would occasionally spike to more extreme errors, exacerbated by the compounding error caused by the use of previous erroneous predictions to make further predictions. The spikes in error rate were reduced in size as the network was retrained, suggesting that this issue may reduce over longer time periods. Figure 7-8 illustrates how the average error varied depending on how many timesteps into the future were predicted: as expected, the further prediction horizons show a higher average error.

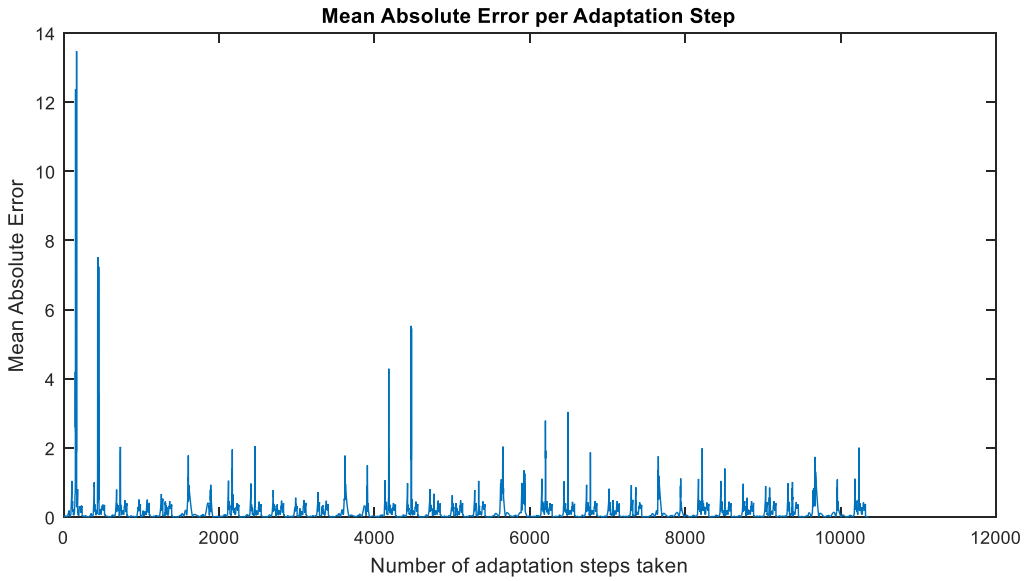


Figure 7-7 – Error Rates of the NARX structure, iterative 12-step prediction across 50-day training period

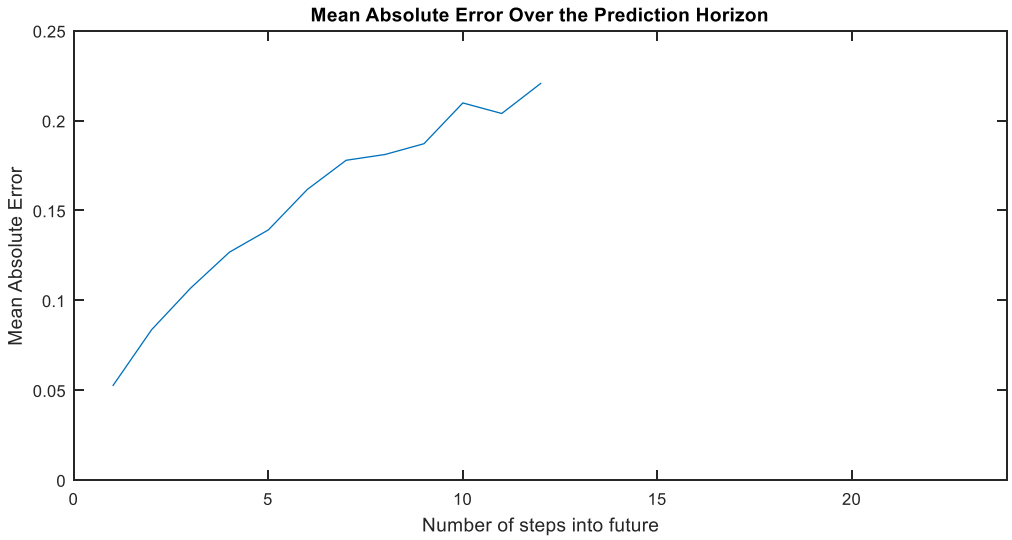


Figure 7-8 - Mean Absolute Error per Prediction Horizon for iterative 12-step prediction

7.6.2 Non-recurrent Network

As a point of comparison, a non-recurrent network was trained on the same data. A structure to predict 12 future timesteps at once was used: similar to that described in section 7.2.1, with the Matlab structural diagram shown in Figure 7-9. The 'feedback' of the past targets was produced manually and appended to the input matrix. This means that the feedback targets and the future predictions were not each directly linked to their own set of time attributes. The only time attributes input were those for the current time.

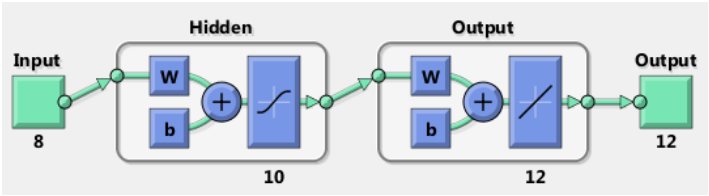


Figure 7-9 – Multiple-Output Simple NN structure on the prediction training set, still trained with trainbr

Over a 50-day run in Figure 7-10 it can be seen that this network structure was less prone to extreme spikes in its prediction error, with all errors averaging to within 2 people of the actual number of occupants, rather than the 14 people seen by the iterative NARX structure within its first week online. However the range of errors was not improved as the network was retrained over time, while the NARX error improved, leaving the average error after 50 days slightly higher on the non-recurrent network. This point is further supported by the average error per prediction horizon

shown in Figure 7-11, which shows no average improvement even on the furthest prediction horizon versus the iterative NARX structure.

The training times, however, were much higher for the multi-output model. Given that the increase in training time with training set size is nearly linear (Figure 7-12) it was decided that the non-recurrent multi-output approach was not a viable option for long-term online training.

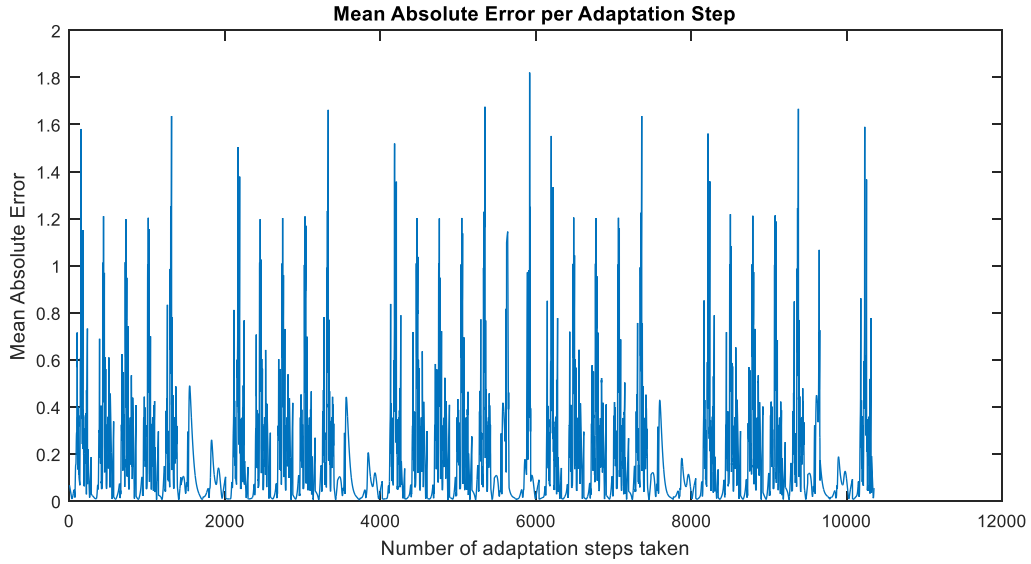


Figure 7-10 – Error Rates of the Multi-Output Nonlinear Input-Output structure across 50-day training period

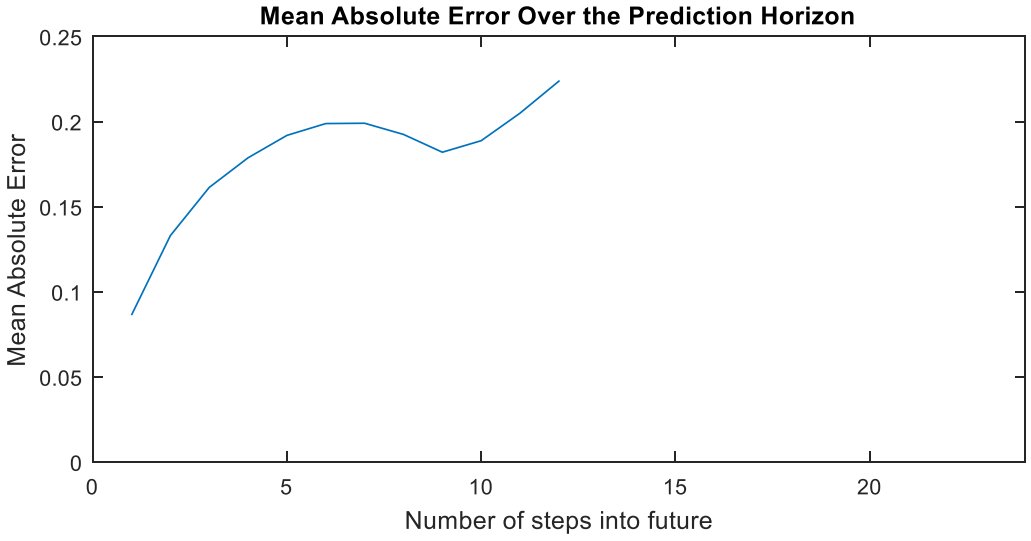


Figure 7-11 – Error Rates of the Multi-Output Nonlinear Input-Output structure across 50-day training period

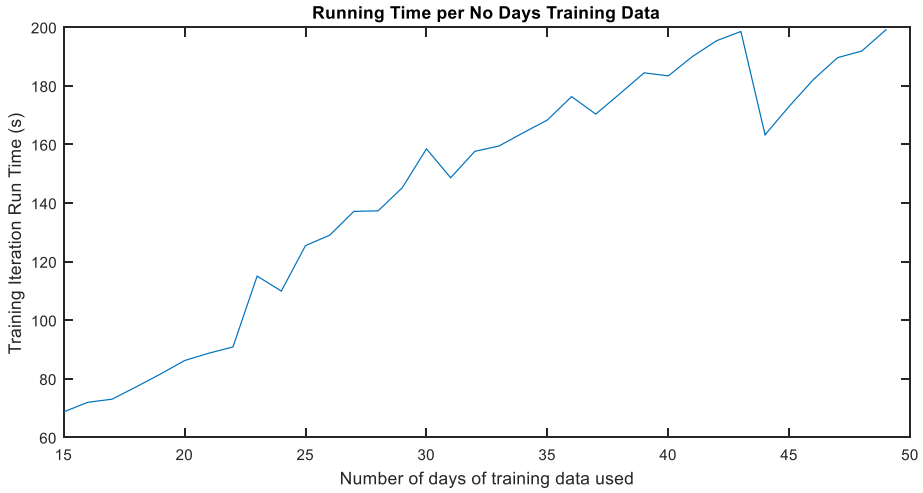


Figure 7-12 – Training Cycle running time vs amount of training data

7.6.3 NARX Network with Multiple Outputs

Figure 7-13 shows a potential NARX structure to predict 12 timesteps into the future in a single calculation, rather than using the standard closed NARX structure to iteratively predict the next 1 step 12 times.

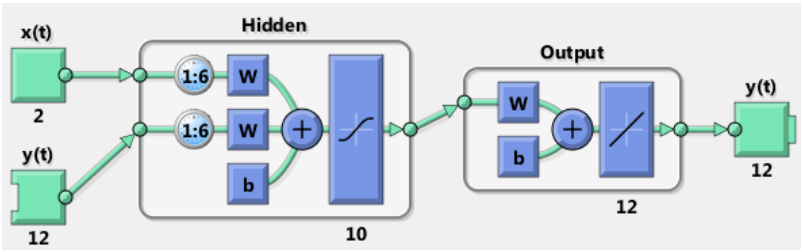


Figure 7-13 - Example of multiple-output open NARX structure

The major issue with this approach was the significantly increased training time and memory requirements. For a single zone, it was not possible to train a model of this structure for 24 timesteps into the future, as the available computer did not have sufficient memory for the task. The initial batch training time for a 12-timestep structure was 1560 seconds, as opposed to 218 seconds for non-recurrent structure. Due to memory constraints on the test computer, it was not possible to run the full 50-day retraining period on this structure, showing that it was not suitable for wider application.

Figure 7-14 also illustrates a potential issue with the fit of the model – the model fit on the initial training data is characteristic of extreme overfitting, with constant oscillation around the training data.

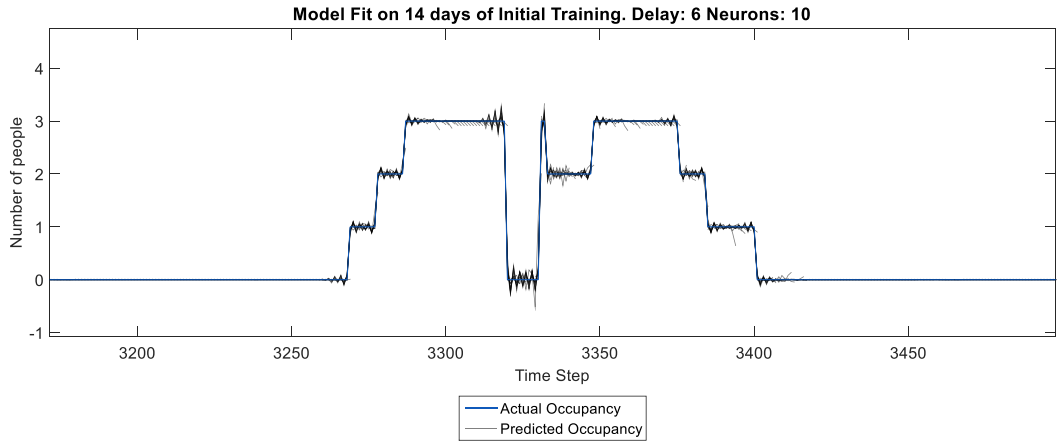


Figure 7-14 – Examples of Model fit on Multiple-Output Open NARX Structure

7.6.4 Selected Base Structure for Further Testing

The structure selected for the initial tests of this model was the iterative NARX method, as indicated in Figure 7-15. As with the initial testing of the detection model, a relatively low number of hidden units was selected to avoid the more extreme overfitting while other factors were tuned. The model was implemented with the manually adjusted Matlab Bayesian method described in Sections 7.3.4 and 7.4. However, in order to cut down computational times and avoid crowding graphs, the final steps to calculate and plot the confidence levels have been omitted in some test stages in the following sections. The prediction horizon was extended to two hours in 24 5-minute timesteps in order to better test the limits of the developed model. In application, the prediction horizon of most interest would be related to the time it would take to condition a space to its desired set point from its current state, which could vary significantly depending on the systems, space and current state in place. A two-hour horizon was considered a balance between feasible prediction and a realistic conditioning time for the Mark Group House application.

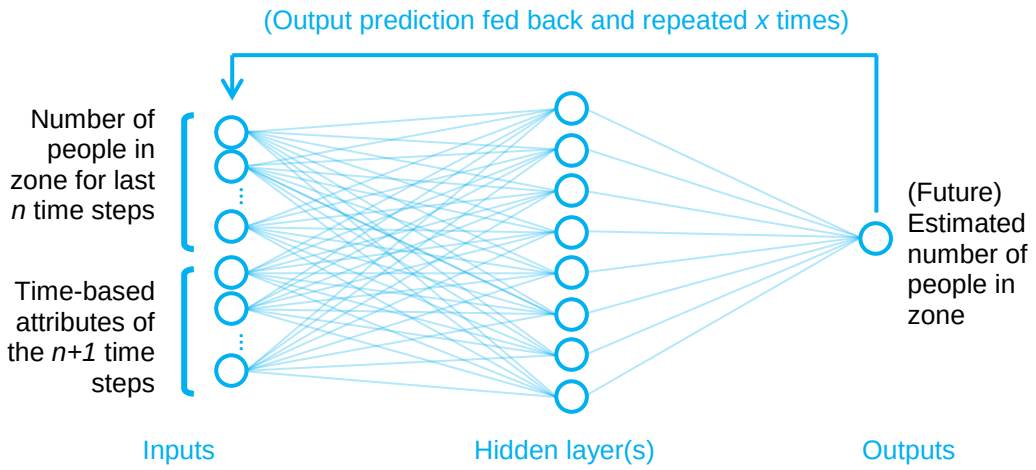


Figure 7-15 – Proposed Prediction NARX Neural Network Model Structure

Detailed testing was planned in order to find:

- Method and frequency of updates to the model using new training data
- Set of features that gives the best performance
    - Time-based features – time of day, day of week, day of year etc.
    - Occupancy features – occupancy rates, binary presence data
    - Number/spacing of previous timesteps included
- Investigation into effectiveness in predicting more or less complex occupancy pattern types
- Optimisation of model structure for reasonable performance/training times
    - Magnitude of error on training/running data
    - Length of training period required to reach acceptable error levels

Throughout the testing conducted, one of the major considerations was achieving an acceptable error performance within a reasonable length of training time. If a system structure could achieve a more complex model of occupancy patterns at the cost of requiring two years of training data before any reasonable predictions could be made, for example, this was considered unacceptable performance. In a practical system, it is not viable to expect such a long time before the system becomes useful.

7.7 Method of Continuous Retraining

One of the major aims for the predictive model was the ability to train 'online': retraining over time in between predictions. This would mean that the model could adapt to change in the underlying occupancy patterns over the long term. Several

different approaches to this aim were considered and tested for their effectiveness, computational requirements and feasibility in terms of time taken.

7.7.1 Matlab 'adapt' function

The Matlab ANN toolbox contains functionality for online model training – the adapt function [225]. This function takes one gradient descent step of the weight/bias values each time it is called. This is technically not a Bayesian training process. However, it should in theory produce a similar optimum result and so was treated as described in section 7.4.

Initial tests of the adapt function on the generated data set showed results that did not meet the expected performance standards. Figure 7-16 shows the MSE over time for two tested networks based on the same training data:

- 'Adapt Only' – network started at a random initialisation, with 'adapt' function called at each time step. In theory, this should start with a high error, which should decrease over time.
- 'Train then Adapt' – network pre-trained in a single batch with 15 days of data prior to the adaptation start. The 'adapt' function was then called at each time step as above. In theory, this should start at a lower error than the random start and either maintain or decrease error over time.

For reference, the MSE of the initial training from 'Train then Adapt' is included in the graph as 'Train Only'. After 500 adaptations of these networks, it was found that both methods using the 'adapt' function performed significantly worse than the standard ANN training process, with the MSE appearing to stabilise around 3-4 for both methods, while the initial training showed an error of 0.13. The reasons for this poor performance were investigated.

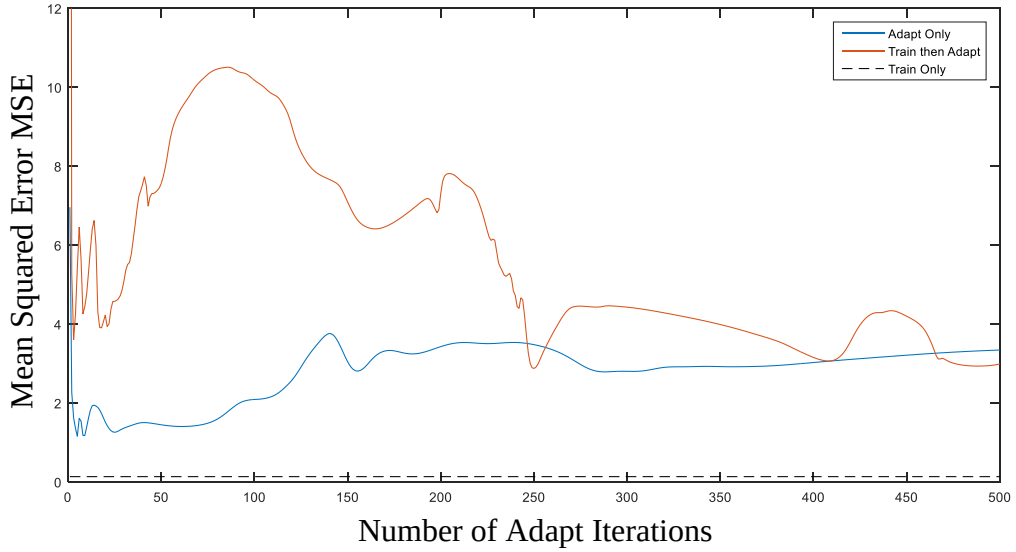


Figure 7-16 - MSE at each adaptation iteration using the same input data at 500 adapt steps

7.7.2 Reducing Adaptation Step Size

A possible explanation for the poor performance of the adapt at each timestep was that the adaptation step size used by the function was too large, causing the model to overshoot its optimum weight values and potentially increase error. The step size was manually adjusted to a range of sizes and compared.

For the Matlab 'adapt' function, the step size at each adaptation is determined by the learning rate *lr* and the momentum constant *mc*. A smaller value of *lr* equates to a smaller step size, which should reduce the average error seen over adaptations if overshooting the optimum weight values is the issue seen in section 7.7.1. The momentum constant *mc* relates to the momentum of the gradient descent, as described in the Matlab documentation [226]. Set between 0 and 1, a higher value gives the adaptation more momentum, allowing it to overcome local minima in the cost function, but at the risk of skipping over the global minimum if set too high. By default, *lr* was set to 0.01 and *mc* was set to 0.09.

In trials of different values for *mc* between 0 and 1, it was found that the momentum had no discernible effect on the error increase seen during adaptation. The learning rate did show some relationship to the error rate, as shown in Table 7-3. It can be seen that the smaller step sizes did give a smaller mean error. However, all of the adapt step sizes still produced a significantly higher error than the unadapted initial training solution.

Table 7-3 - Overall RMSE from adapting with different Learning Rates

LR 0.0001	LR 0.001	LR 0.01	LR 0.1
2.135	2.325	2.292	1.124e+152

Figure 7-17 shows that the mean error visibly increased the more adaptations occurred. From the graph, it appears that the error from each learning rate converged to a value around 2.3, with the lower learning rate taking longer to reach this average, while the higher learning rate showed more variation once it reached this average. As each learning rate separately approached this increased error value, the source of the increased error could not be from a mismatch of the learning rate causing overshoot of the optimum solution. Further investigation was required.

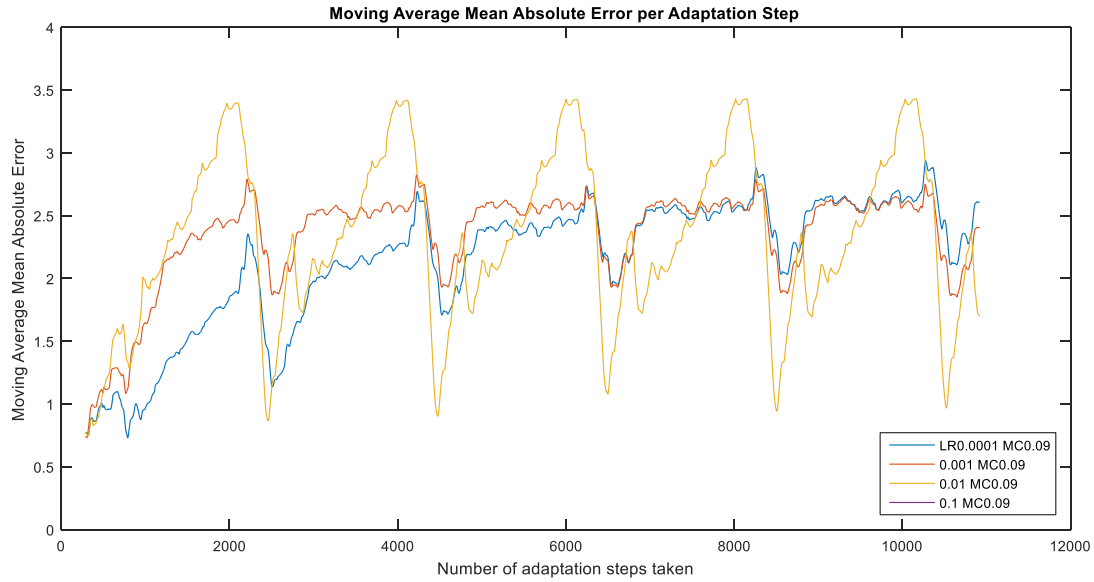


Figure 7-17 - Moving Average Mean Error per Adaptation Step for Varying Step Size

7.7.3 Adapting every Timestep – further investigation

The poor performance of the adapt function was further investigated by plotting the output from the network at each timestep. Figure 7-18 shows one day of the NARX network after an initial batch training of a 9-day dataset. It can be seen that the network correctly predicted most general trends in occupancy through the day, although difficulty is encountered around the short lunchtime drop to 0 occupants. Figure 7-19 shows the aggregated results of one day's adaptation, having started on a 9-day initial training from the same simulated dataset as Figure 7-18. Here, it can be seen that the network begins to favour prediction of a constant occupancy rate, significantly reducing the performance of the prediction relative to the non-adapting

case. This is counter to the intention of adaptation. Some potential conclusions to draw are:

- The long periods where occupancy stays constant outnumber the periods when occupancy varies, due to the nature of an intermittently occupied office space.
- This appears to bias the adapting network towards solutions that provide the simplest explanation for the data seen – occupancy stays constant
- As the adaptation updates its weights by only one step per timestep, the times with varying occupancy are not sufficient to 'pull back' from the training obtained during long constant-occupancy periods, and such steps are quickly reversed during constant occupancy again.
- This may be able to be manipulated by adaptation step size, but the solution more likely to avoid this problem is to adapt in larger batches, perhaps overnight at the end of each day or week.
- This finding verified the existence of an optimum between responsiveness to changing circumstances and maintaining a robust model.

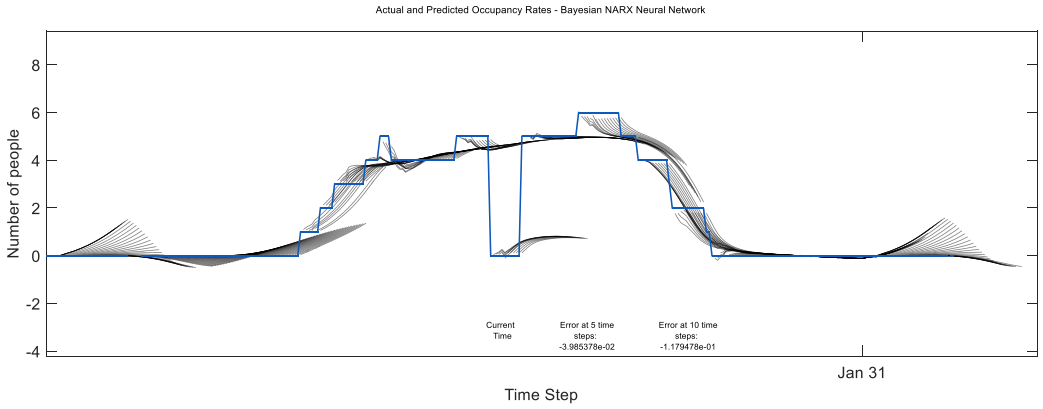


Figure 7-18 - Non-Adapting NARX run on Simulated Data

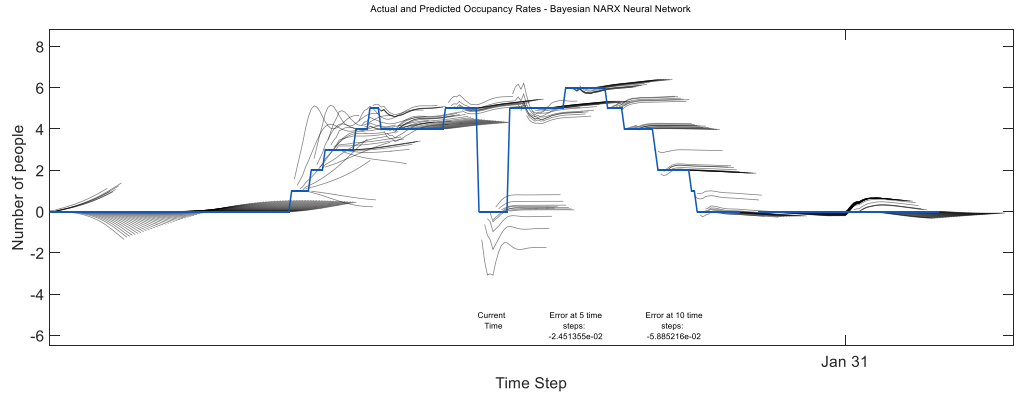


Figure 7-19 - Adapting per Time Step NARX run on Simulated Data

7.7.4 Adapting in 1-day Batches

As an alternative, it was tested whether calling the adapt function at the end of each day would improve the results shown. It was found that the 1-day adaptation suffered from the same issues as adapting at each timestep, as shown in Figure 7-20.

Upon inspection of the adaptation process used by Matlab, it was found that calling adapt on a batch of new data points runs iteratively through each new point, suggesting that the adaption process may simply not be suitable for this application. The long unoccupied periods during the night cause a steady state solution to be favoured too strongly to be countered by the daytime values.

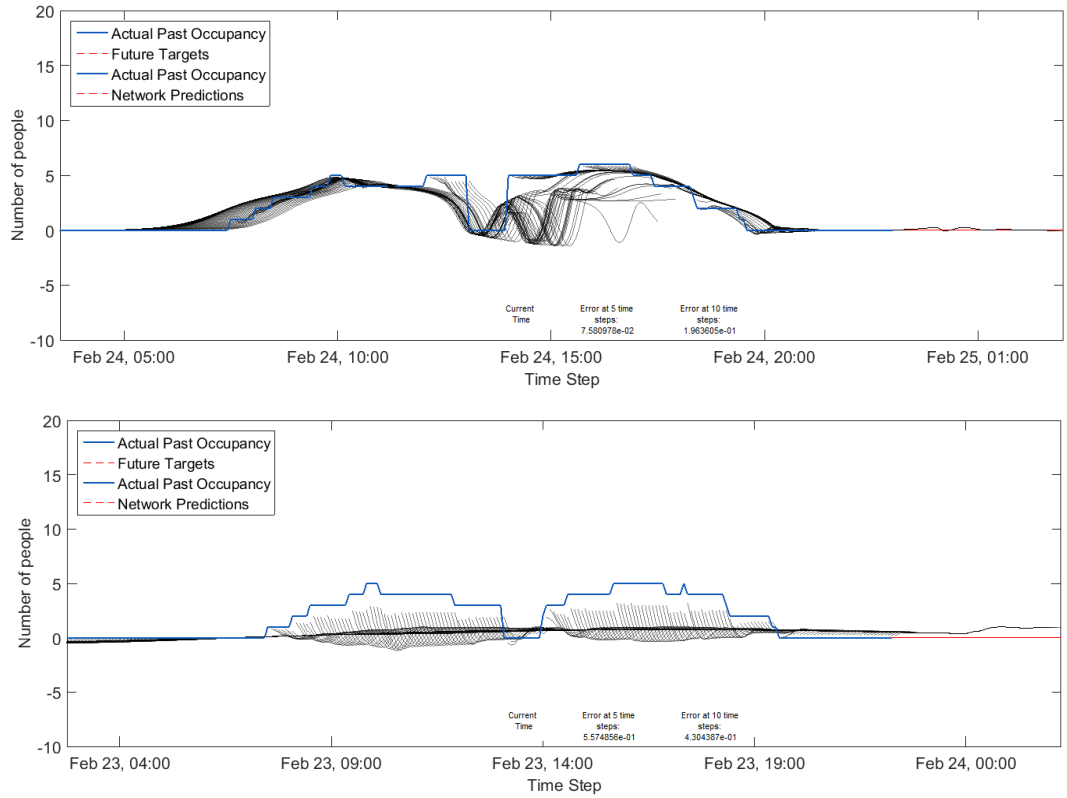


Figure 7-20 – 1-day Batch Adapt – Sample of a) Initial Training b) Model after 1 day of Adapt Data

7.7.5 Batch Full Retraining

Another potential option was to fully re-train the network periodically during its online operation. This process accounts for all new data points at the same time and so should not suffer the bias towards constant values that was seen in the single-step adapt function. However, by the same process it was expected to need a significantly longer computational time than the ‘adapt’ function, and so was not feasible to update

the network on every timestep. At each retraining, the training dataset consisted of the full initial train dataset, plus all data points up to the current time from the 'online' dataset. In order to limit the increase in training time as more and more training data was included, a one-year 'sliding window' was used on the total length of the training data, meaning that after a year of training, older data would begin to be discarded in favour of new data.

Figure 7-21 shows a sample from the end of a 100-day run of daily retraining on a generated dataset. In comparison to the equivalent from the inbuilt 'adapt' function, as seen in the previous section, this shows a much more successful fit to the data after the adaptation process. However, it was found that the training time required to completely retrain the network each day was high enough to inhibit repeated testing for the later stages of optimisation.

It was investigated what effects would be seen if the frequency of retraining was decreased. 100-day runs were made on the same base dataset, retraining every 1, 2, 7 and 28 days. The moving average RMSE for each of these runs is shown in Figure 7-22. It was found that the error rate for most frequencies decreased within the first 4 weeks of retraining before stabilising around a similar level. The one-day retraining frequency saw several days with a much higher spike in error rate, likely caused by poor instances of retraining on particular days, where the cost function reached a less effective local minimum during training. While the lower frequency trainings did not encounter a particularly bad retraining during the 100-day run sampled, it should be noted that if a bad retrain was to occur, it would take longer to be corrected if the model was updated less frequently. It was therefore decided to remain with a daily retraining frequency.

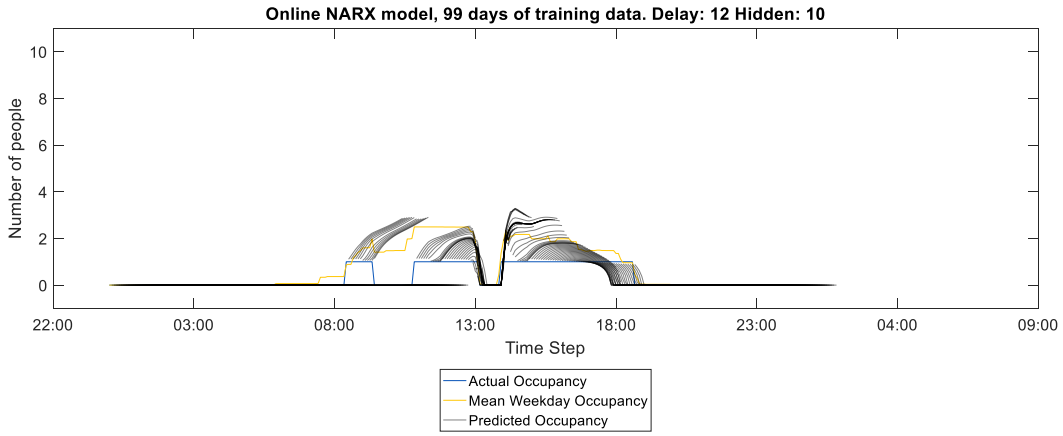


Figure 7-21 – Sample of model output after 100 day Daily Full Retraining

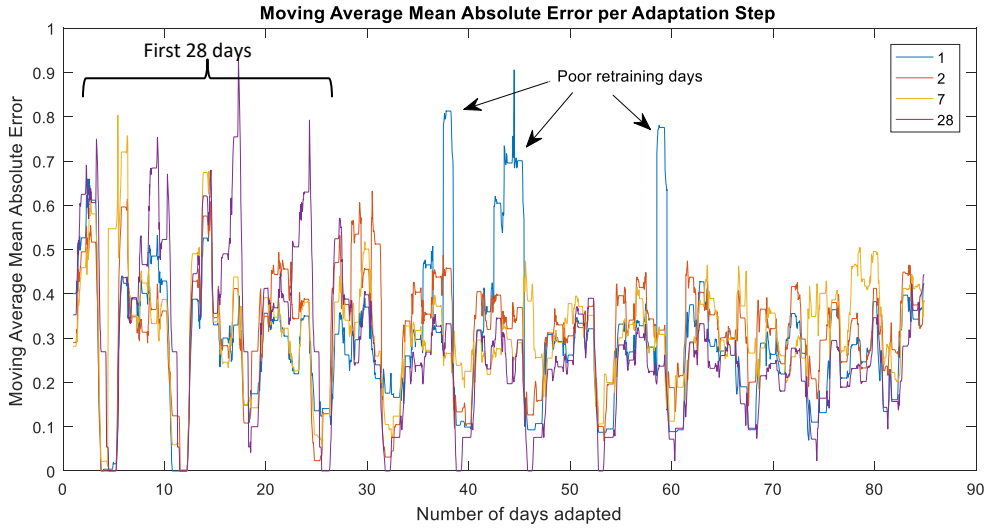


Figure 7-22 - Moving Average Mean Abs Error per Time Step – Varying Frequency of trainbr retraining, 100 day run

Parallelisation of the training process

Once work began on the realistically-sized training set, it became clear that the training process was prohibitively slow to quickly calculate the effects of changing various options while adapting over time. One measure to decrease the calculation time was to parallelise the training function. In MATLAB, this can be achieved relatively easily, as MATLAB has inbuilt parallel calculation functionality. However, it was necessary to re-edit the trainbr function to ensure that the parallel-mode functions also output the Hessian and other parameters used to calculate the error bars.

To illustrate the reduced calculation time, the same model structure was trained on the same 50-day online training set (14-day initial training, 36-day online training). The run times were as follows:

	Single Worker	**MATLAB Parallel Workers**	**Percentage reduction**
Total Run Time	791.504 s	617.795 s	22%
Training-specific functions run time	667.458 s	473.021 s	29%

Parallelised training reduced the training time by almost 30%. This translated to a 22% reduction in the total run time for a short training period.

7.8 Feature Selection

7.8.1 Time-based External Input Variables

Given that the occupancy data contained patterns on a range of time scales, it was investigated how much value was provided to a network of the same structure trained on 100 days of training data when including following time-based data as an external input:

- Day of Week – represented as an integer value 1-7. This allows the day to be input as a single variable, but encodes some continuity between days, so that Monday (day '1') is perceived as more similar to Tuesday (day '2') than Sunday (day '7'), where in reality it is one day away from each. This could produce undesired effects in the trained network.
- Day of Week, binary – represented as seven discrete inputs with a 0/1 value. This avoids the issues discussed above, but the larger number of inputs could introduce overfitting issues.
- Weekday/Weekend – represented as a single 0/1 value to denote if the day is a weekday or weekend day.
- Time of Day – represented as an integer 1-288 to denote each of the 288 five-minute slots in a day. This representation suffers some of the issues described above, but is not viable to include as 288 separate inputs.
- Day of Year – represented as an integer 1-366.

Figure 7-23 shows a comparison of the mean RMSE over a 2-hour prediction horizon across the 85 days of runtime after an initial 15-day training period, where a higher RMSE indicates a worse performance. It can be seen that the binary day of the week gave a poor performance, while combinations featuring the time of day produced the lowest errors. The following analysis was made to pick out some of the ways in which these features affected the performance in specific situations.

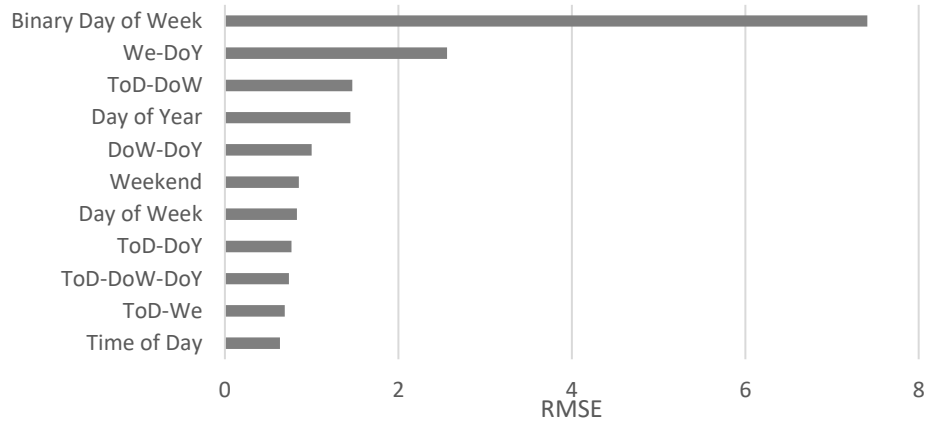


Figure 7-23 - RMSE Comparison with Varying Time-Based External Inputs

The difference in model output caused by the input format of the weekday is illustrated in Figure 7-24, which shows the actual training data versus the model's representation when trained on this data. When provided with the binary day of the week, the model shows a nonzero value on the two weekdays at the end of the training set, despite the actual occupancy being (unusually for the typical occupancy pattern) zero. This is likely indicative of a closer association with certain patterns being strongly connected to particular days. However, the model was clearly prone to overfitting issues, as illustrated in Figure 7-25, which shows the model output fluctuating around extreme values when the model is presented with new data from the 100th day of the test dataset.

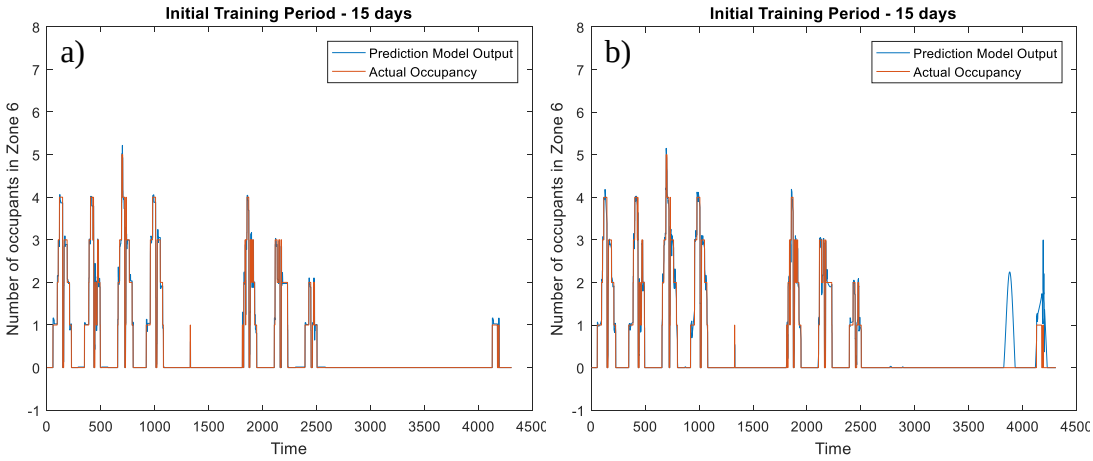


Figure 7-24 - Training Data vs Model Representation for a) Integer and b) Binary Day of Week Models

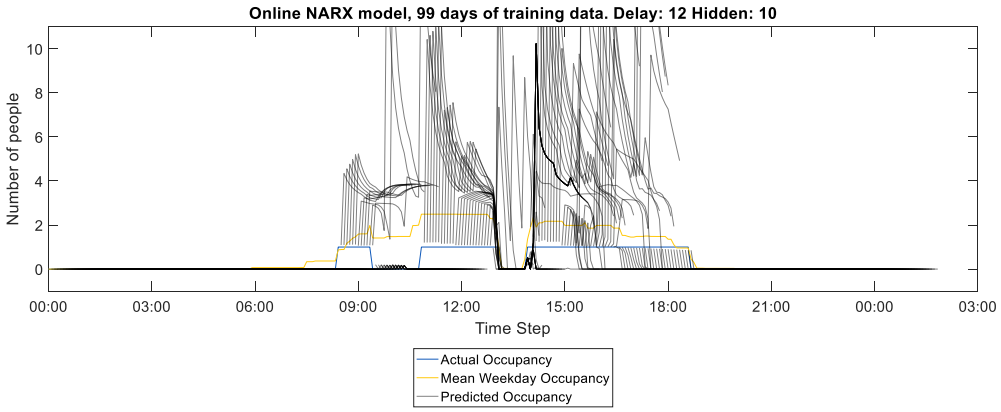


Figure 7-25 - Sample of Model Output on 100th day of training, Binary Day of Week Input

While the error rate alone shows that the time of day data was beneficial, Figure 7-26 illustrates how the inclusion of this data changed the model output. Graph a) shows that when provided weekend data only, the model always predicts a similar pattern of rising occupancy when the zone had been recently vacated. This is because the model has no way to distinguish between the short absence seen at lunch break and other absences. When provided with the time of day as in graph b), the model can then define the strong pattern of a short absence at midday without predicting the same return of occupants when all occupants leave in the evening.

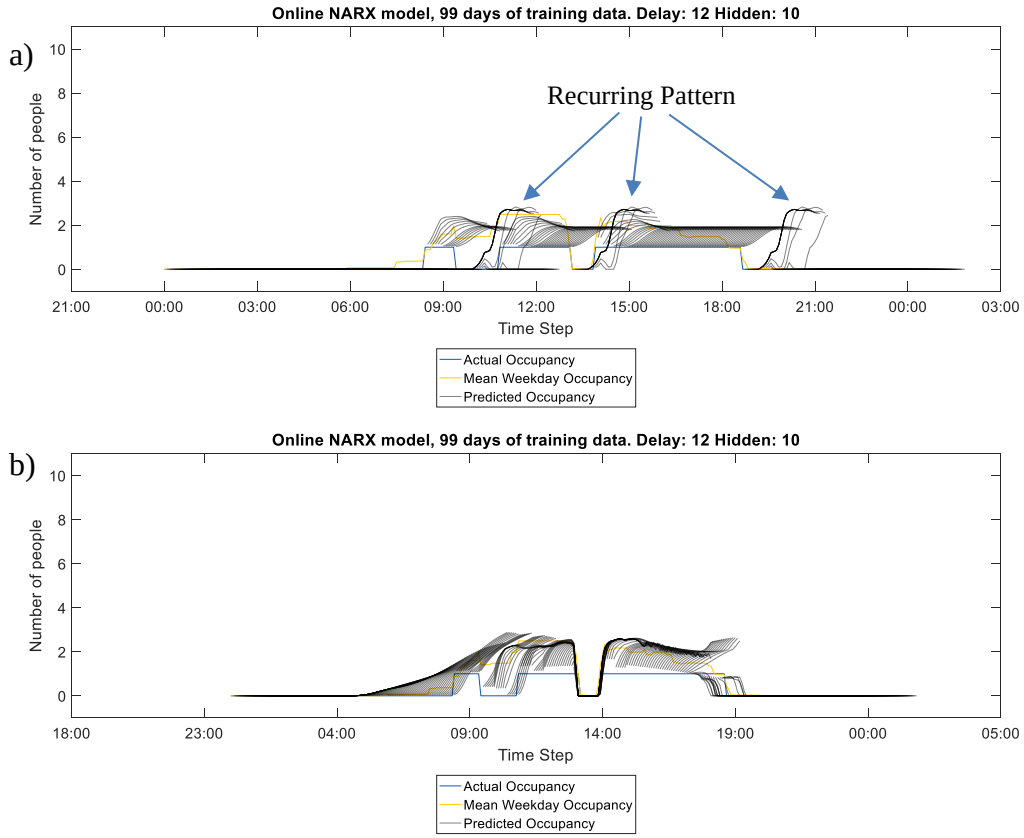


Figure 7-26 - Sample of Model Output on 100th day of training for a) Weekend Data Only b) Time of Day & Weekend Data Input

A final observation was made on the difference in the model's behaviour on unusual weekdays when supplied with some data combinations. It was found over multiple random-initialisation trainings that the Time of Day-Day of Week combination tended towards a fit more biased towards the current observed occupancy level than other combinations, as shown in Figure 7-27, where both graphs show the same day of the test data. The model trained on the time of day, day of week and day of year tended towards predictions that occupancy would rise to its typical weekday levels, while the model trained on just time of day and day of week was more responsive to unusual situations. The reason for this was not clear: none of the time-based attributes should specifically allow for these unusual situations. It was assumed that this effect was seen due to random variations in the fit from different trainings and may not have occurred had a much larger number of trials of each data combination been conducted. However, this effect was considered in the further studies of model structure below.

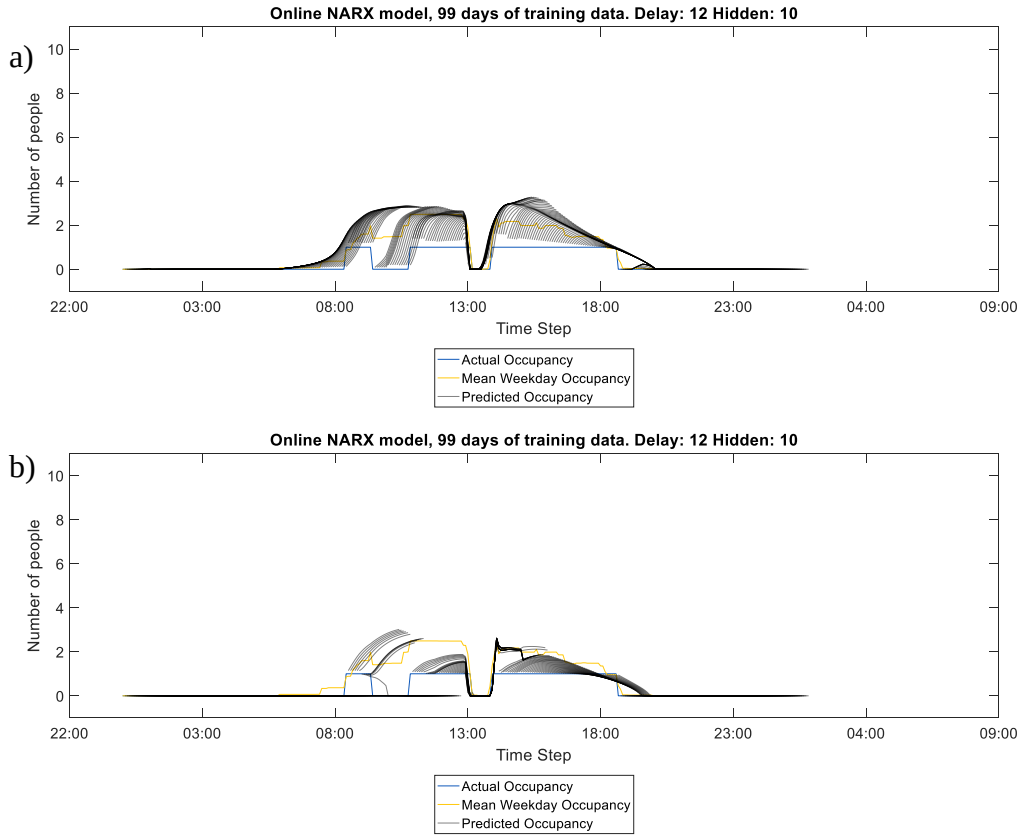


Figure 7-27 - Sample of Model Output on 100th day of training for a) ToD-DoW-DoY b) ToD-DoW Input

7.8.2 Including binary presence of occupants in zone

In the occupancy data provided for training, the general trends in local occupancy rate are informed by individual patterns from each occupant agent. It was investigated whether including specific data on known occupants would improve the performance of the prediction model. The past presence of each individual known occupant was included as 16 discrete binary inputs to the network.

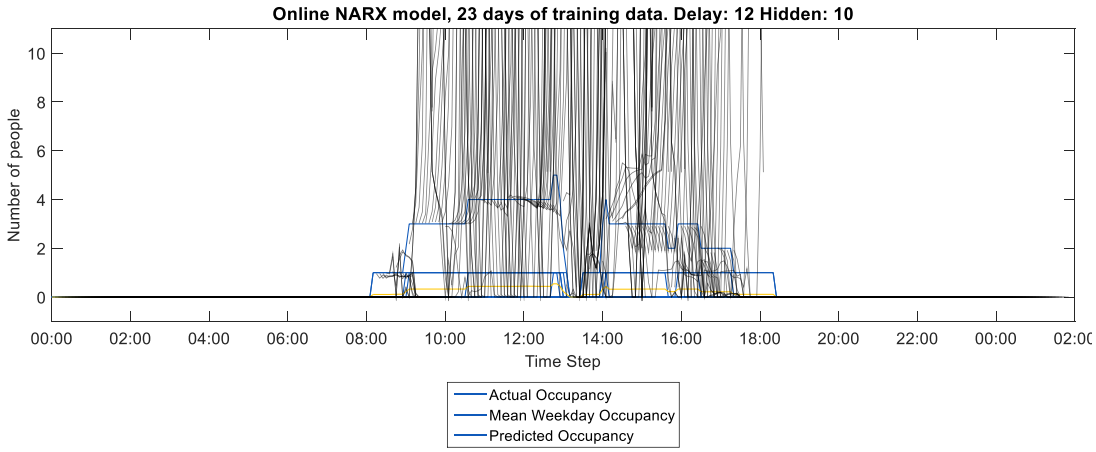


Figure 7-28 - Sample of Model Output on 24th day of training, binary presence of individual known occupants included

Figure 7-28 shows a sample from the training of a network trained with this additional binary presence input. The resulting model is clearly prone to extreme fluctuations, likely indicating overfitting issues. Providing a larger initial training dataset could reduce this issue, but it was deemed unacceptable that the model would not reach anything close to stable operation within a 3-week setup period.

The overfit problem could be potentially reduced by including fewer occupants, such as only those occupants who are known to spend the majority of their time in the zone being trained. However, Table 7-4 shows that both presence-data-inclusive models performed significantly worse than an equivalent model without these inputs. The error was slightly reduced by including only the most relevant individuals, but no by a margin large enough to justify including this data at all.

Table 7-4 - Comparison of Average RMSE of Prediction Model using Occupant Presence Data

	No individuals	All individuals	Select individuals
RMSE	0.250	1.730	1.540

7.8.3 Number/Spacing of previous timesteps included

As it was previously shown that the model had a tendency towards overfitting to the training data, it was important to ensure that the total number of inputs, and so internal weights, of the model was not excessively high. As each previous timestep included is treated as a separate input, it was investigated how far back into the past it was necessary to include for adequate training of the model. In the following tests, the 'delay', or number of previous timesteps included in the network training was varied on several training runs using the same 100-day generated dataset.

One of the behaviours of the predictive model that had been observed in previous runs was the tendency to predict a typical number of occupants returning from a lunch break, regardless of the number of occupants seen on a morning. This led to an overestimate on unusual days when fewer occupants were present, as shown in Figure 7-29. It was tested whether including further previous timesteps would reduce this issue. However, it was found that as the number of previous steps was increased, overfitting issues leading to unstable predictions on new data were increasingly strong, as shown in Figure 7-30. It was clear that the total number of previous timesteps included could not extend beyond 12 individual inputs.

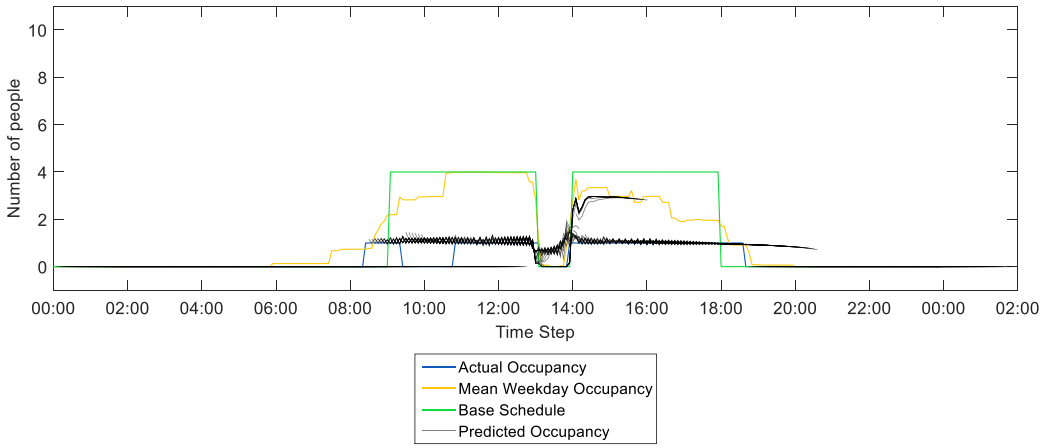


Figure 7-29 - Sample Unusual Day - Illustration of Lunch Break Prediction when supplied with up to 12 (1 hour) previous timesteps

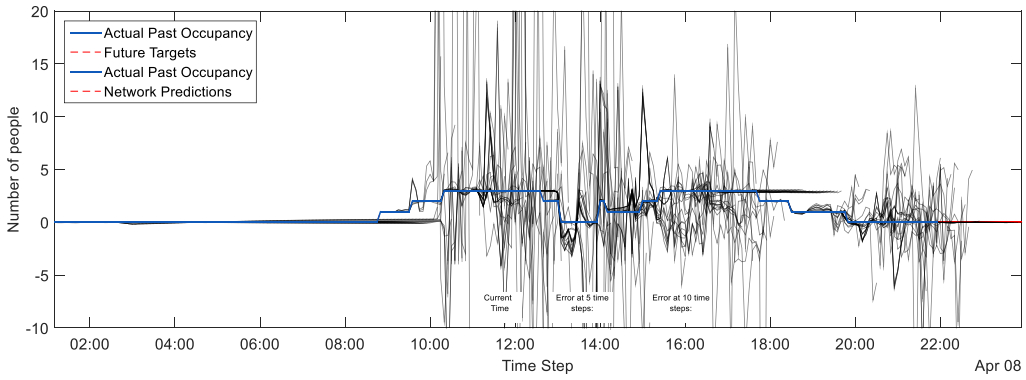


Figure 7-30 - Sample Day when supplied with 15 (1.25 hour) previous timesteps – overfitting encountered

A series of tests were made with reduced delays. It was also tested whether non-consecutive timesteps could be an effective way to reduce the number of individual inputs while maintaining the furthest horizon included: for example, a delay of '1-7-12' takes three past timesteps in total, spaced 5, 35 and 60 minutes into the past respectively. The 1-day moving average RMSE of some of the more successful delay variations is shown in Figure 7-31. It can be seen that the RMSE was generally reduced by reducing the number of previous timesteps included as model inputs. Where the same number of inputs were used with different timeslots, the slots further into the past showed a higher general error rate, suggesting that these inputs were less relevant to the future occupancy.

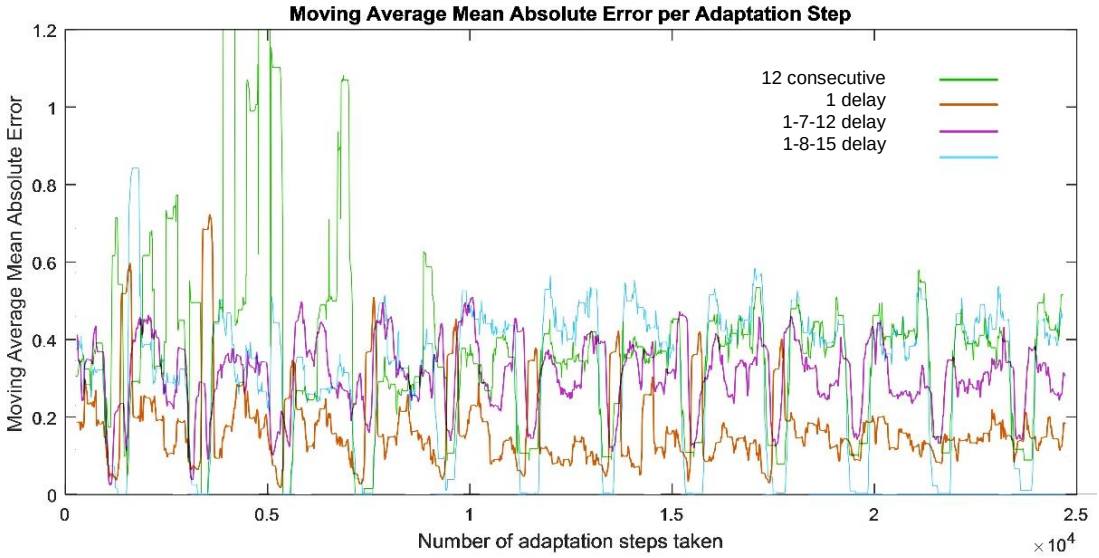


Figure 7-31 - Comparison of Varying no Previous Timesteps - RMSE over 100-day run

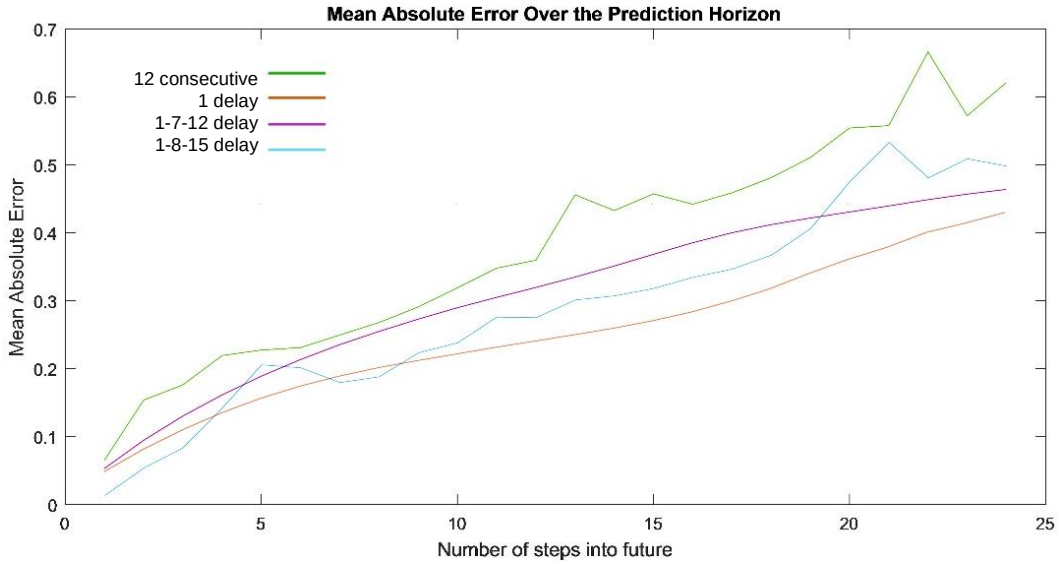


Figure 7-32 - Comparison of Varying no Previous Timesteps – average RMSE over prediction horizon

7.9 Recognition of pattern types

The generated dataset was used to assess how the complexity of the underlying occupancy patterns affected how useful the prediction model was relative to simpler prediction heuristics. The two simpler rules tested were:

- Base – this 'prediction' was based on a typical unchanging occupancy schedule, as might be found applied to a standard building control system. It assumed that all occupants arrived at 09:00 and left at 18:00, with a 1-hour break 13:00-14:00. Weekends were assumed to have zero occupancy.

- Mean – assuming that accurate occupancy data is being collected, a simpler way to predict occupancy for a given time of day on a given day of the week would be to take the mean of all previous observed occupancy rates at that time and day.

It was hypothesised that the simpler heuristics would perform less well against the prediction model the more realistically complex the observed occupancy patterns became.

7.9.1 Regular patterns only

The model was initially trained on a dataset generated with occupant agents that never deviated from their personal default schedules, as defined in Section 7.5. This meant that the occupancy of each zone followed a set weekly pattern with no variations. From Figure 7-33 and Figure 7-34 it can be seen that the prediction model was prone to instabilities on this highly regular dataset, giving a relatively high error rate. The 'mean' method of prediction performed perfectly with this unrealistically regular occupancy data due to the lack of variation on the standard weekly schedule.

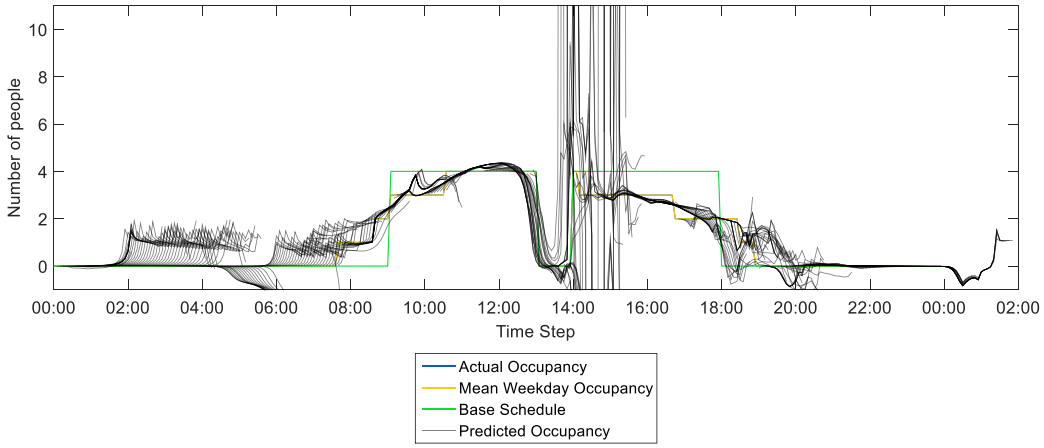


Figure 7-33 - Regular Occupancy Only - Sample of multi-occupant zone after 99 days of training

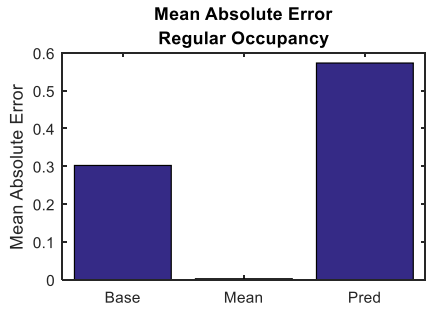


Figure 7-34 - Regular Occupancy Only – Comparison of Error Rates for each Prediction Method

7.9.2 Regular Patterns + Gaussian Noise on start/end times

Building up the realistic variations on the agent occupancy profiles, the next training was run on a dataset with the same regular patterns as previously, but with a Gaussian-based agent-dependent variation on each occupant's entry/exit times for any occupied period. Figure 7-35 shows a sample day at the end of a 100-day training period. The instabilities in the prediction model are no longer present, with a smoother prediction less prone to extreme fluctuations. The mean prediction had become less effective.

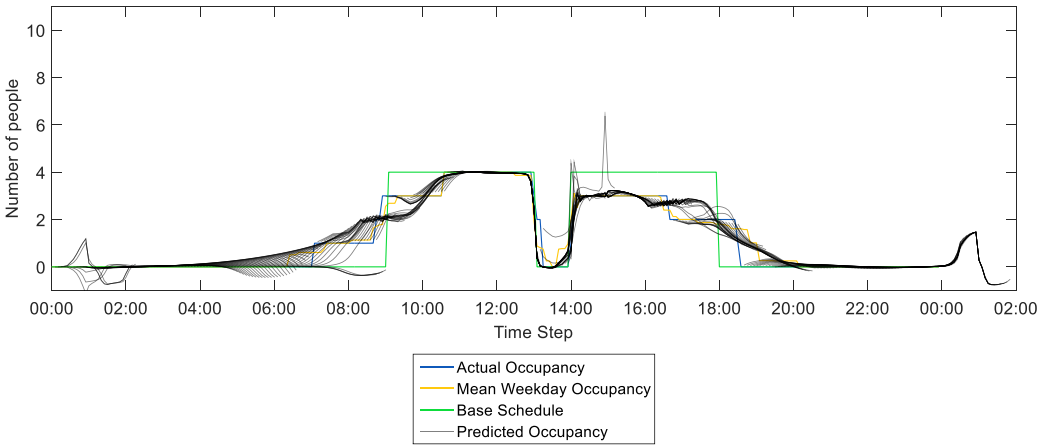


Figure 7-35 - Regular + Noise - Sample after 99 days of training

7.9.3 Noise and Random Events

The next iteration included additional random events that took occupants out of their regular zones for short periods of the day in order to simulate non-regular meetings and other typical daily events. Figure 7-36 shows the prediction model sample after training, which continues the trend of the mean and base schedules becoming less effective for prediction purposes.

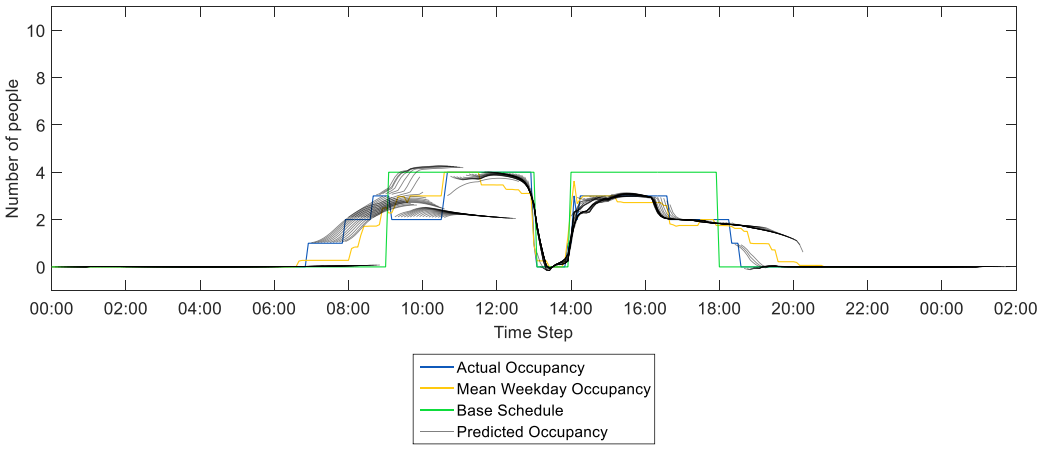


Figure 7-36 - Regular + Noise + Random Events - Sample after 99 days of training

7.9.4 Noise, Random and Visit Events

The next addition to the dataset complexity was visit events, where non-regular visits to occupied zones were included, simulating meeting attendees, students attending drop-in sessions with staff, site visit groups etc. Figure 7-37 shows the prediction model sample after training, which further continues the trend of the mean and base schedules becoming less effective and the trained prediction model producing more realistic estimations based on the current occupancy.

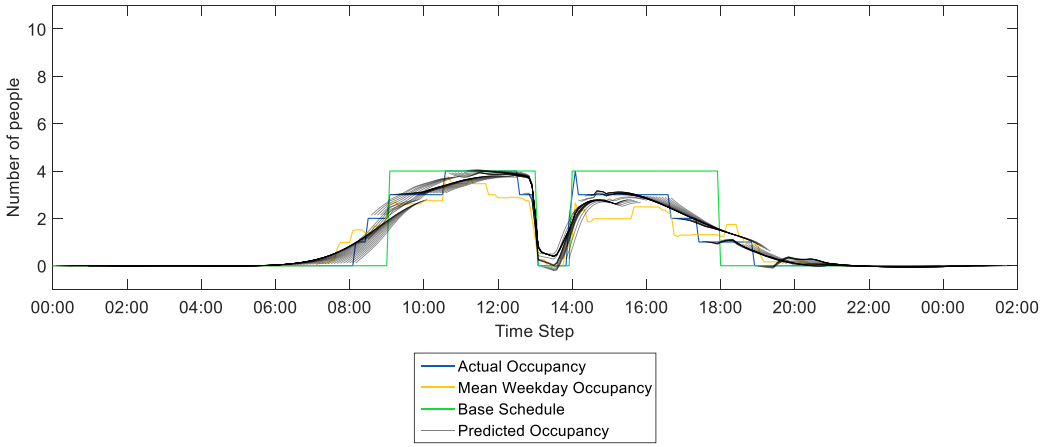


Figure 7-37 - Regular + Noise, Random, Visit Events - Sample after 99 days of training

7.9.5 Full Complexity - Noise, Random, Visit and Holiday Events

The final dataset complexity included all the factors listed in Section 7.5, with the same base schedules as the previous examples and including the time-of-year based holiday events where agents would be absent for a prolonged period around certain times of year. In this case, at times close to national holidays when most office occupants are on holiday, the daily number of occupants in a zone can differ significantly from the usual profile. In the case of this training data, one of these unusually low-occupancy days was found at the end of the 100-day test period, as shown in Figure 7-38. Here, the trained prediction model outperformed the simpler heuristics by a much larger margin, as the model was able to react to an unusually low number of occupants logged in the recent past and adjust its future predictions accordingly. Figure 7-39 shows a more typical weekday for comparison.

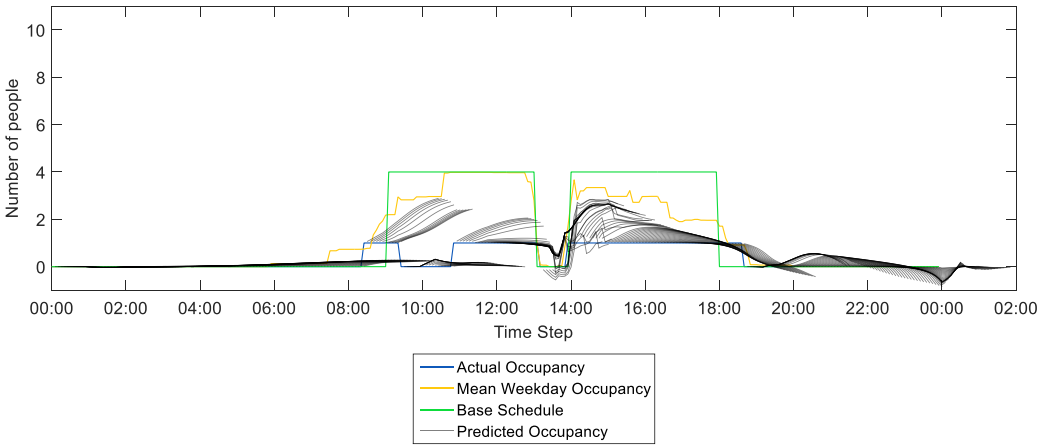


Figure 7-38 – Full Complexity Simulated Dataset – Unusual Weekday Sample

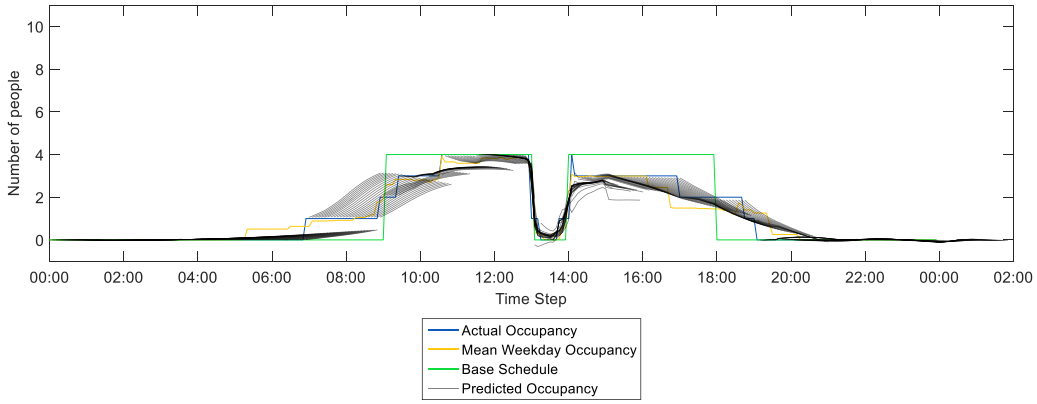


Figure 7-39 - Full Complexity Simulated Dataset – Typical Weekday Sample

7.9.6 Full Complexity – Positive-Constrained Model

It is possible to constrain the output from an ANN model to only non-negative values, however this was found to increase some overfitting issues found in Chapter 6 and so was not tested as the default structure for the prediction model. However, as the full-complexity prediction model appeared to show less signs of overfitting than the detection model, it was tested with a positive-constrained output to check if this reduced the error rate by ensuring that negative occupancy values could not be predicted.

Figure 7-40 and Figure 7-41 shows this model's results on the same reference days to Figure 7-38 and Figure 7-39, respectively. It can be seen that the positive-constrained model produced predictions that stayed tighter to the actual occupancy rate, but with a much stronger, and sometimes inappropriate, bias towards the current value of occupancy, as shown in the unusual day, where the model fails to predict that the present occupant will leave until after they have done so.

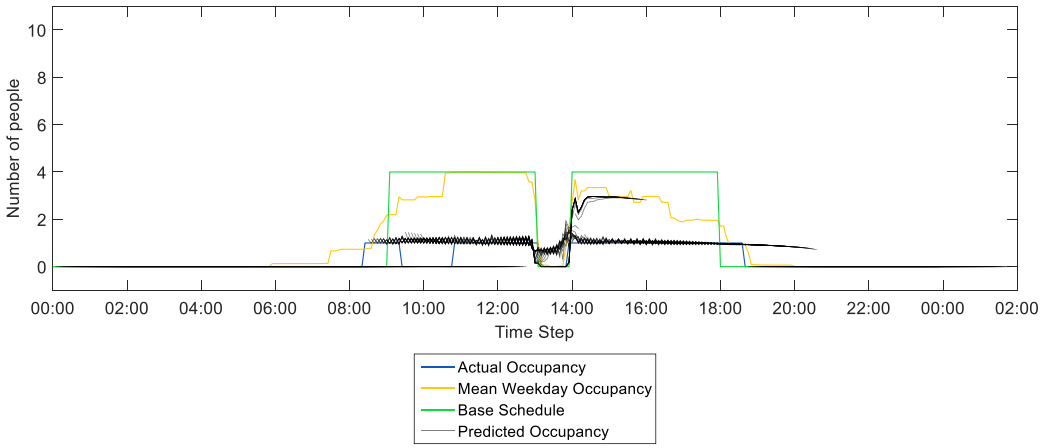


Figure 7-40 - Full Complexity Simulated Dataset, Positive Output - Unusual Weekday Sample

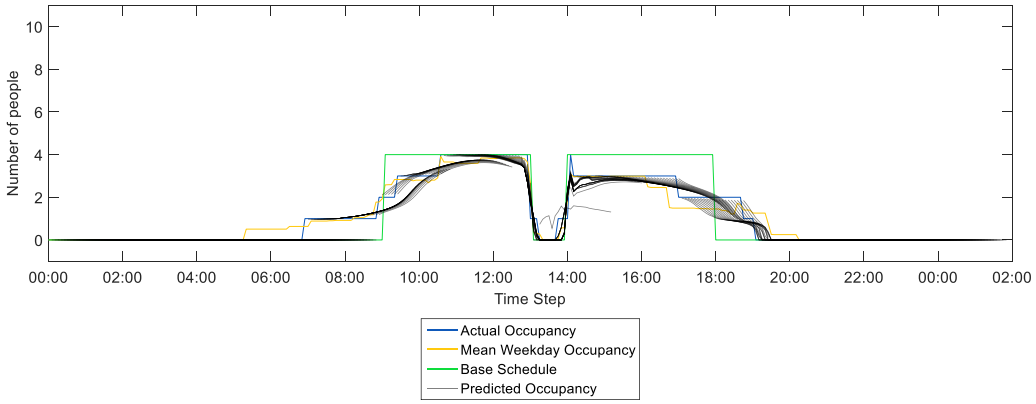


Figure 7-41 - Full Complexity Simulated Dataset, Positive Output - Typical Weekday Sample

7.9.7 Comparison

One measure to assess the effectiveness of the trained prediction model was to calculate how far into the future it was able to produce more accurate predictions than the other two methods. Figure 7-42 shows the comparison to the Base Schedule for each dataset complexity level: as the dataset became more realistically complex, the trained prediction model was able to outperform the baseline up to 110 minutes into the future on average. When compared to the Mean Schedule in Figure 7-43, it was found that the trained model showed a less pronounced, but still positive improvement, outperforming the mean up to 50 minutes into the future on average. A clear relationship can be seen in both cases, that the trained model was able to account for unusual circumstances where the simpler methods could not.

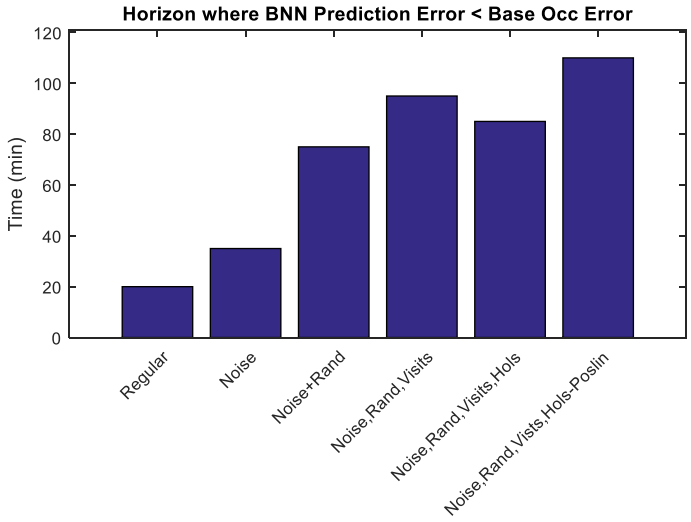


Figure 7-42 – Horizon with lower RMSE of Trained Prediction Models against Base Schedule

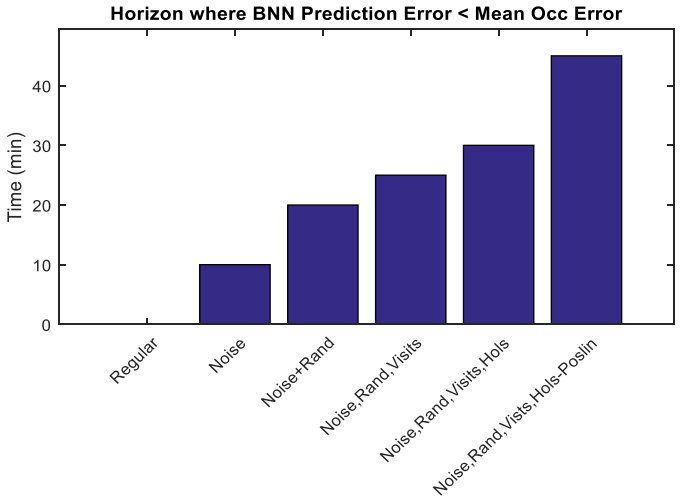


Figure 7-43 - Horizon with lower RMSE of Trained Prediction Models against Mean Schedule

It was also tested whether the outputs from the prediction model could accurately predict binary presence – whether the tested zone was occupied by any people or not. For the chosen regression model, presence was assumed when the prediction was greater than 0.5. Occupant absence was assumed when the presence was less than or equal to 0.5. Once again, this accuracy was compared to simpler prediction methods, including the binary presence assumed by whether the running daily mean was greater or smaller than 0.5.

Figure 7-44 shows a sample of the binary accuracy of the predictive model when compared to the mean. Where the bar is above the axis, the model is outperforming the mean method and when the bar is under, the mean is outperforming the model. Figure 7-45 shows a summary of the performance across all data pattern types. From this analysis, it was possible to observe the following:

- The accuracy of the mean estimation was independent of the time horizon, as it was not based on the current reading.
- This meant that for shorter prediction horizons, the BNN prediction produced more accurate predictions.
- As a general rule, the more randomness/complexity measures were included in the dataset, the longer the prediction horizon where BNN was more effective than mean.
- With the model setup tested, for the most realistic dataset, BNN produced a better accuracy than the mean for an average of 40 minutes into the future.
- For most of the datasets, the prediction performed better on weekends.
- On the holiday-inclusive datasets, the common holiday periods could be clearly identified by the performance of BNN prediction over the mean prediction. BNN performed consistently better during these periods, as indicated in Figure 7-44.

It should be noted that this analysis does not account for the uncertainty level obtained in the BNN prediction. In application to a control system, the uncertainty could be used to supplement the final prediction, potentially allowing some of the erroneous predictions to be identified and ignored in real time. This could potentially further improve the performance of this option.

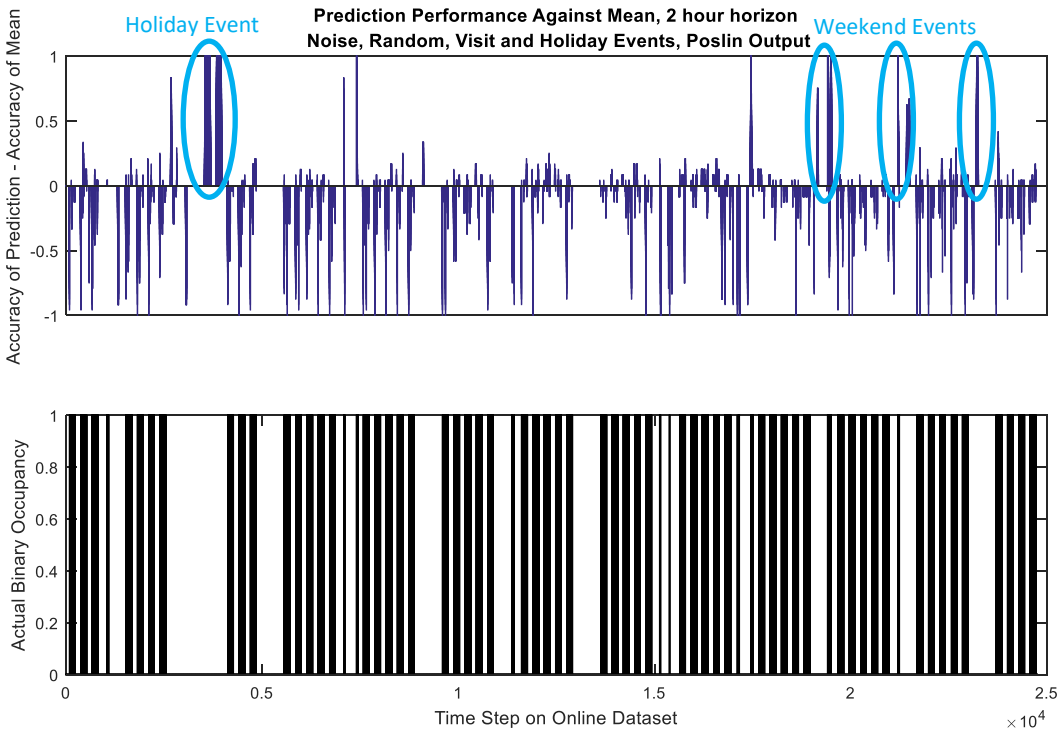


Figure 7-44 – a) Comparison of Binary Accuracy from Predictive Model vs Mean Occupancy b) Binary Occupancy During this Period

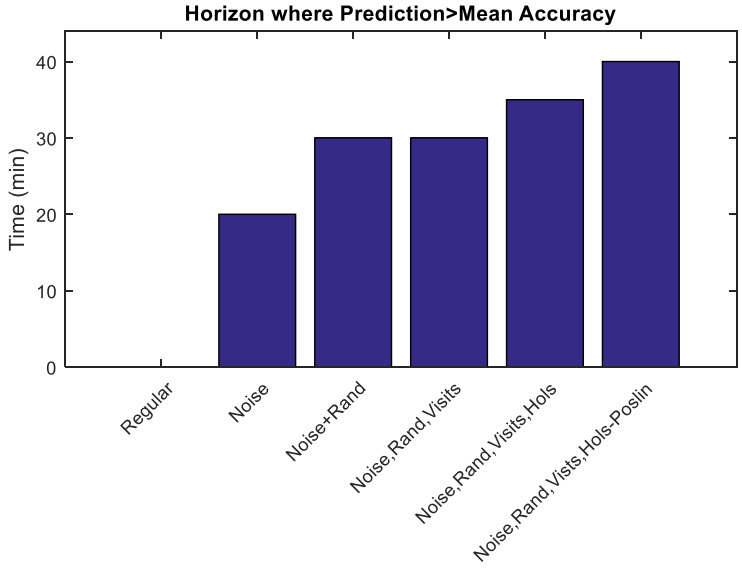


Figure 7-45 – Binary Presence - Horizon with higher accuracy of Trained Prediction Models against Mean Schedule

7.9.8 Long-term Changes in Pattern

In order to test a long-term change of occupant agent, a dataset was generated with a change of one agent in a multi-occupant space at around 2/3 of the full dataset length. Figure 7-46 shows the results of this test. The error rate over time showed no

significant increase in the average error after the occupancy change was introduced. This is likely because:

- In a multi-occupancy room, the relative effect of a single person changing patterns is low
- Given the multiple occupants, but no way to define between them, the model in this form does not capture features detailed enough to be greatly affected by a single occupant change

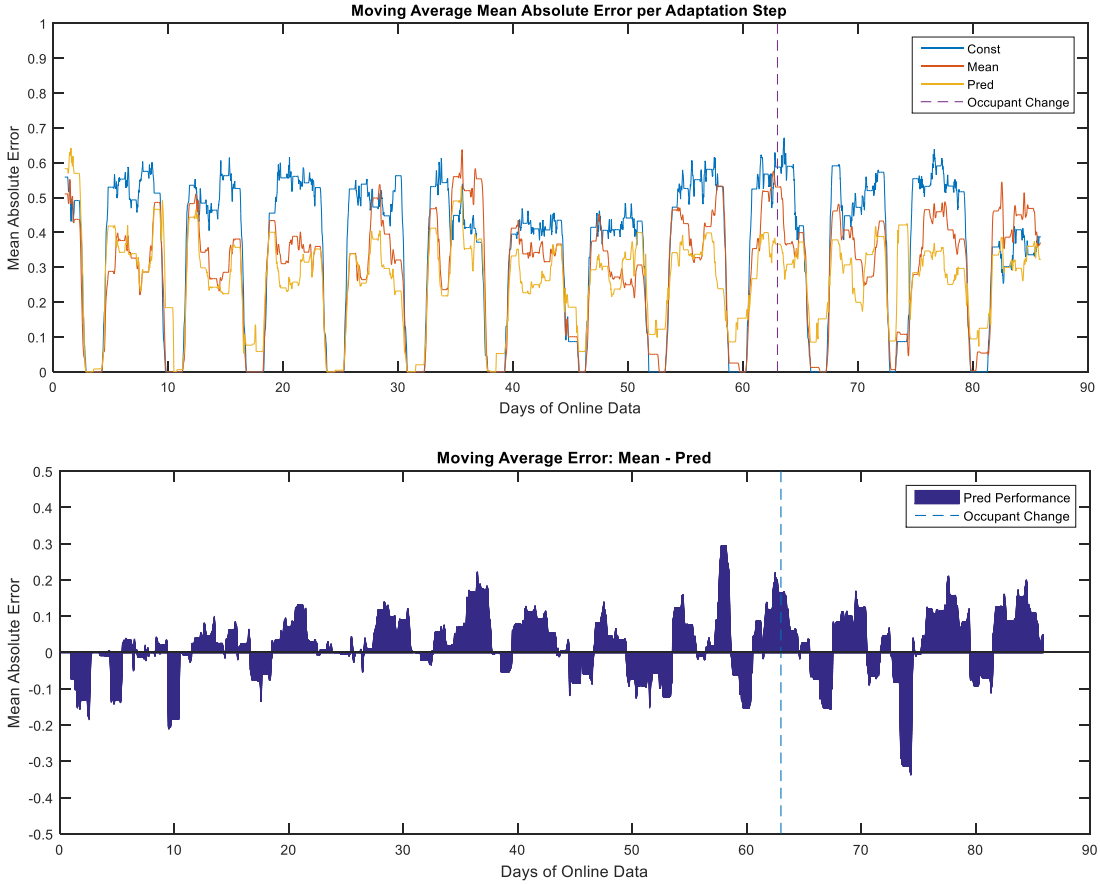


Figure 7-46 - Mean Error of Predictive Model vs Mean Method after Long-term Change in Agent a) Mean Absolute Error b) Comparison of methods

In terms of performance against the mean strategy, the results were not clear. It does appear that the prediction model performed on average better than the mean after the occupancy change, but this was often also true before the change. It should also be noted that the worst performance against the mean occurs after the occupancy change – on a weekend when the post-lunchtime increase in occupancy was falsely predicted by the model.

7.10 Model Optimisation

7.10.1 Deep Learning

Although the initial tests of the model to define the most useful sets of inputs etc. were conducted on a relatively simple network structure with a single hidden layer, it was investigated whether adding further layers would allow for the capture of more complex behaviours in the occupancy patterns. A comparison of several multi-layer configurations was made, although due to computing restraints the number of neurons in the hidden layers was required to be kept to a similar level as the simple testing. As the system was repeatedly shown to be prone to overfitting even at low numbers of neurons in the previous testing, it was assumed that much more complex deep learning structures would not have been able to reach an acceptable performance level within the 1-4 week boundary set for reasonable operation of the model in a real application.

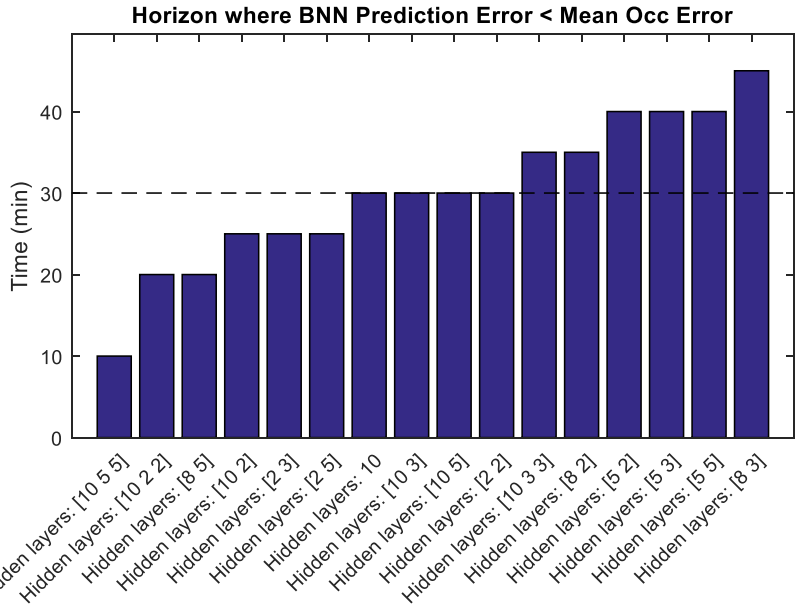


Figure 7-47 - Comparison of Multi-Layer Networks – Max. Prediction Horizon that outperforms Simple Prediction Heuristics

Figure 7-47 shows that the average beneficial performance against single-layer network of 10 neurons for the full-complexity test dataset is varied. The high variation between similar structure types suggests that the training tends to fall into local maxima and so depends on the initial state of the network. This means that the comparison in the graph should not be taken as definite ranking, but as a general indication of what types of structure were more or less successful.

Figure 7-48 and Figure 7-49 show the moving average performance of each structure over time, demonstrating some of the differences between the structure types. The two-layer structures starting with a 2-neuron layer showed a relatively low error on the first 2 weeks of training data, but began to underperform relative to the other structures as the length of training data available increased. The structures with an initial layer greater than 8 neurons showed a much larger error during the first three weeks of adaptation data, and then settle to a performance similar to the other structures over time. This suggests that overfitting is an issue with smaller datasets as the structure becomes even slightly more complex.

In general, it appeared that a two-layer approach can yield a better performance than a single-layer over the long term. For the three-layer models, it appeared that overfitting to the training data became more of an issue. It should also be noted that most of the three-layer structures tested required more than one attempt at training, as they ended towards local minimum solutions such as prediction of zero occupancy at all times.

Another consideration when increasing the complexity of the network is the computational training time required in order to achieve a certain level of accuracy. There are many ways to include this as a metric for success. The total training time over a 100-day run for each variation is shown in Figure 7-50, while Figure 7-51 shows the average error weighted by total training time. Here it can be seen that the structures with fewer neurons offer a better error per runtime. However, runtime can be sacrificed to some extent to benefit the accuracy over the long term, as long as the system could feasibly run in real-time. It was considered that any of the structures up to the three-layer configurations would be feasible to run as part of a system re-training daily.

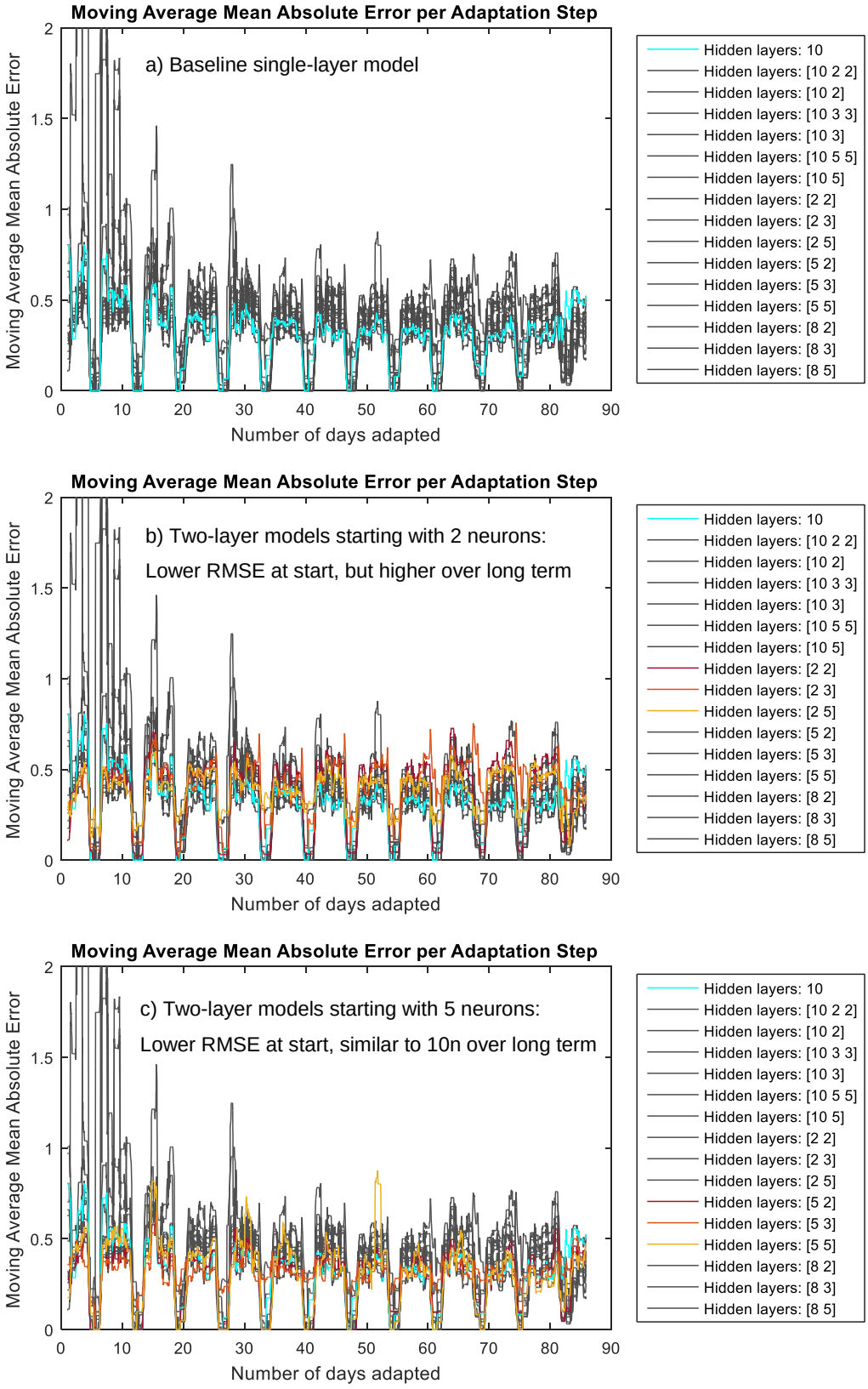


Figure 7-48 a-c - Moving Average Mean Error for various Prediction BNN structures over 100-day Training Period

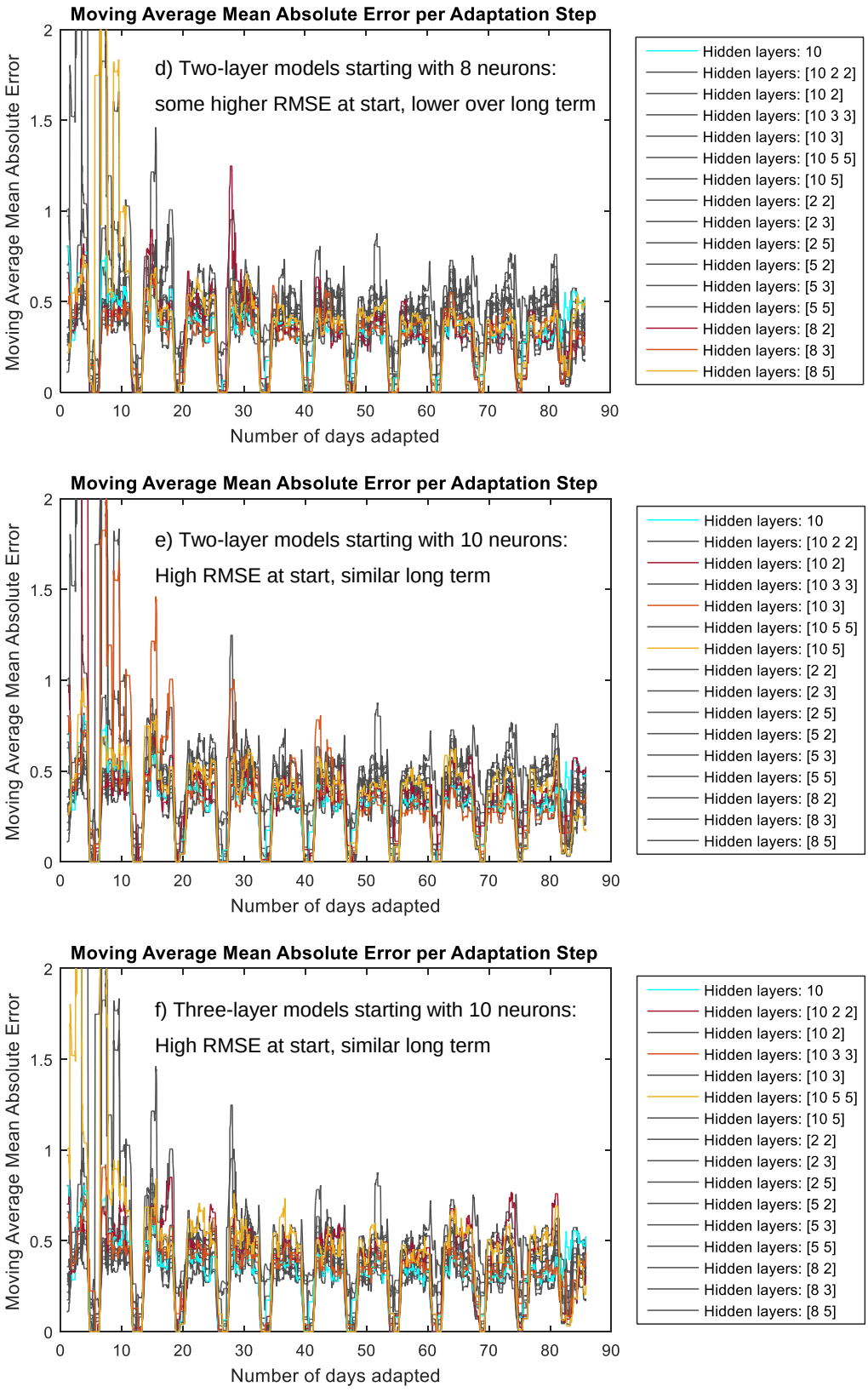


Figure 7-49 d-f - Moving Average Mean Error for various Prediction BNN structures over 100-day Training Period

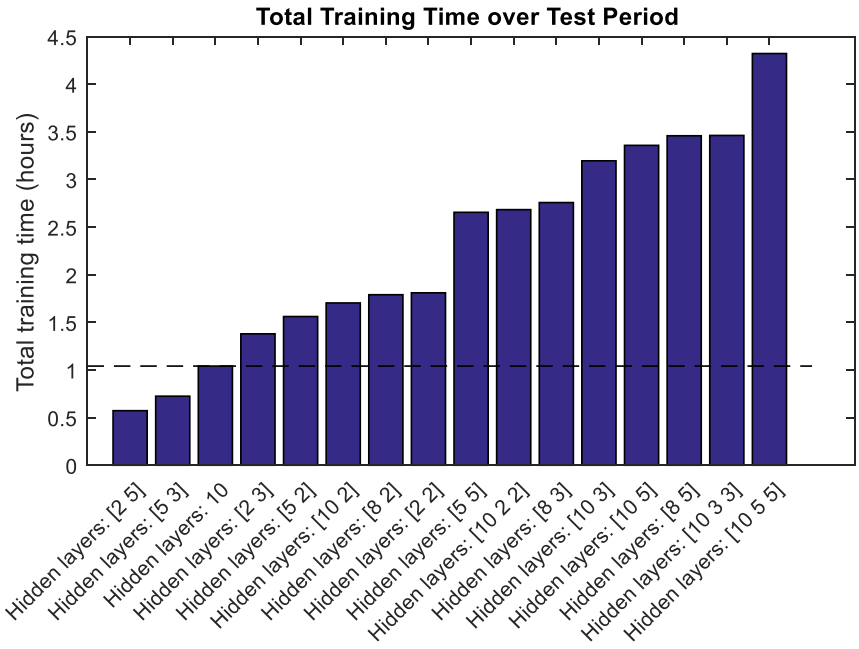


Figure 7-50 – Total training time over the whole 85 day online training for various BNN structures

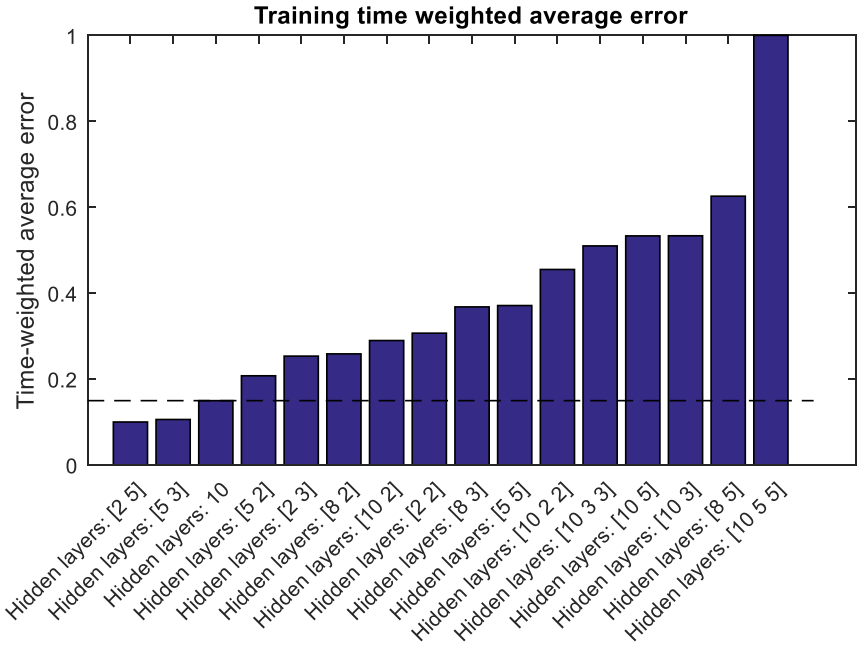


Figure 7-51 - Error averaged over the whole training period, weighted by multiplying by total training time/max training time

7.11 Proposed Model

A finalised model was proposed to combine the optimal findings as discussed in the previous sections. This included the following:

- NARX neural network trained with the Bayesian optimization to allow for calculation of uncertainty levels on the predictive output.
- Daily retraining of the model at the end of each day to update the patterns found with new data as occupant schedules change over time.
- Inclusion of the time of day and day of week as model inputs.
- Superfluous previous timesteps omitted from the model inputs.
- Relatively simple internal structure of hidden layers – as more complex structures were found to overfit badly with smaller amounts of training data, they were considered unsuitable even if they could have had the potential to encode more complex patterns over a much longer timescale of training. An 8-2 hidden layer structure was selected.

It was found in section 7.9 that the trained predictive model could perform better than all tested simpler prediction heuristics up to around 45 minutes into the future, and against the baseline fixed schedule up to 2 hours. It was also observed that the highest errors in the prediction model output were most often associated with a high value of the calculated uncertainty variable. In order to improve the overall performance of the model after the 45-minute prediction horizon, a hybrid model was proposed to defer to the mean-based heuristic when the output of the model was too uncertain.

The optimum threshold for the uncertainty level was determined by testing the overall RMSE found with several different thresholds. It was found that during a 100-day run of the trained prediction model that the uncertainty variable σ^2 showed a wide range of values, indicating an extremely low certainty in the model output at some points. The range of threshold values for σ^2 used are shown in Table 7-5.

Table 7-5 - Hybrid Model Values Tested

Model Name	σ^2 Threshold Value
Hybrid 1	1
Hybrid 2	10
Hybrid 3	100
Hybrid 4	1000
Hybrid 5	10000
Hybrid 6	100000
Hybrid 7	1000000

Figure 7-52 summarises the performance of each of these hybrid models against the fully predictive model and the fixed-schedule baseline as described in section 7.9. Hybrid 1 favours the mean-based heuristic more often, while Hybrid 7 favours the trained model in all but the most uncertain cases. It can be seen that at the lower thresholds, the far-future prediction performance is significantly improved, but the near-future is less accurate, as the hybrid model is predicting the mean-based value more frequently even within a 15-minute prediction horizon. This suggests that a balance can be made depending on the prediction horizon of most interest, which is determined by the system to be controlled and its projected response time in the space being monitored.

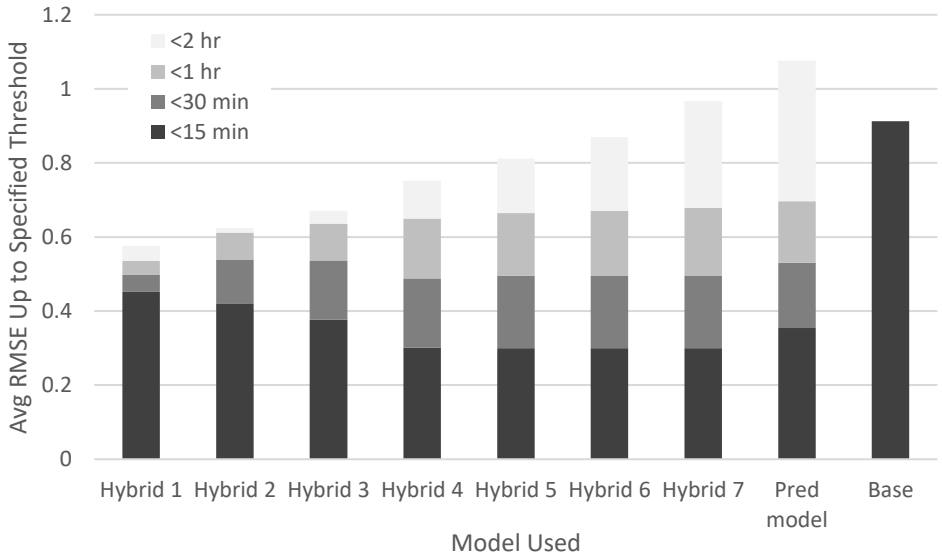


Figure 7-52 – Comparison of overall Hybrid Model Performance against Fully Trained Model and Baseline Schedule

In the case studies made in Chapter 4 and the review of existing research, it was established that times when the occupants' behaviours significantly deviates from typical patterns are the most critical in terms of energy saving opportunities. The balance of performance between the trained prediction model and the mean-based

heuristic in these unusual situations was compared for only days with unusual occupancy patterns. This was defined as any weekend day with nonzero occupancy, or any weekday that had fewer than the expected number of regular occupants for the given zone all day. Figure 7-53 summarises the RMSE for each of the hybrid models for unusual days only: it can be seen that the mean-favouring thresholds performed significantly worse up to a 1-hour horizon during these days, with the optimum overall performance found with Hybrid 5 at a threshold of 10000. This threshold appeared to provide the best balance between ignoring extremely uncertain outputs from the trained model, but allowing for some increase in uncertainty that is seen when an unusual situation arises. It is this threshold that is therefore proposed for the final predictive model.

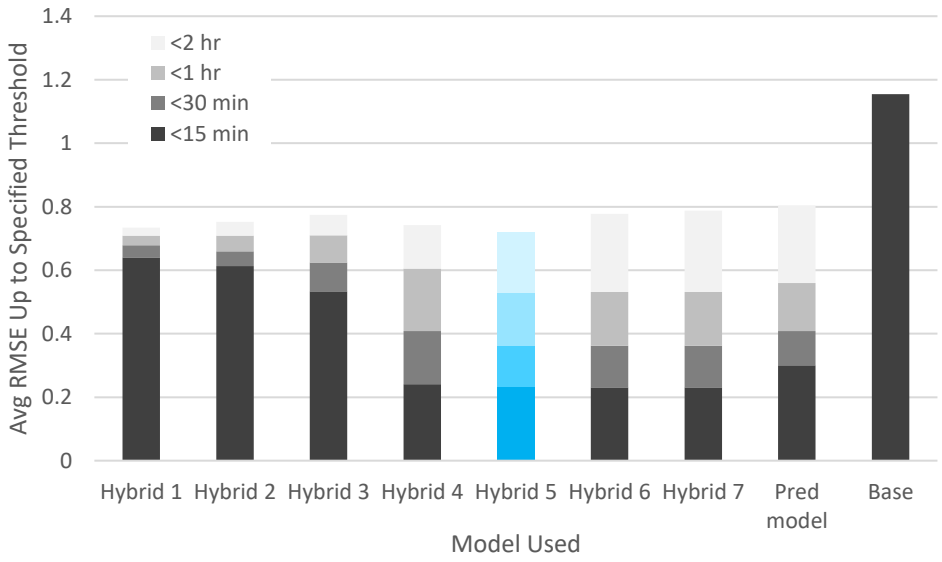


Figure 7-53 – Comparison of overall Hybrid Model Performance against Fully Trained Model and Baseline Schedule over days with Unusual Occupancy Patterns

It should be noted that the exact structure and numerical thresholds discussed above for the final model are location-specific and would likely need to be adjusted if the predictive model was to be applied to a different building. Selection of appropriate values for these parameters of the model could be automated for a more general application, if supplied with an amount of training data and a known optimal prediction horizon/pattern trait to aim for.

From the graphs shown, it can be seen that the final proposed model was able to predict the local number of people in a multi-occupant space with an average error of less than one person, with increasing accuracy at closer prediction horizons. The

output of the model is complemented by a calculated uncertainty level that can be supplied to building controls systems in order to better inform the logic used to optimise building energy use in response to measured and predicted demand levels. A sample of the Matlab code used to construct the final models is included in the physical copy of this work in Appendix 10.7.

7.12 Conclusions

In this chapter, a model to predict the changing number of occupants in a space across the near future was developed based on measured values of recent past occupancy rates. A Bayesian Neural network was selected to allow for the calculation of uncertainty levels in the model output, with the acknowledgement that achieving a perfect prediction of future events is not possible, the level of certainty in the predictions made is a valuable measure for the purpose of making control decisions.

The structure and optimisation of the proposed model was developed using a generated dataset to allow for testing on training data of an arbitrary length and control over the types of patterns shown in the data. A method to generate a dataset with same characteristics as the test building was developed as a part of this work, with a basis on using individual occupant agents with individual variations on characteristics such as typical times of occupancy and likelihood of staying longer than usual. While the use of generated data allowed for more detailed enquiry into the timescales required to reach a relatively stable error rate and the types of pattern that could be predicted by a trained model, it should be noted that there are potential limitations to not testing on long-term sets of real-world occupancy data. For example, the generated dataset may not represent all patterns and complexities found in real-world data. A potential extension to this work would be to apply a similar testing strategy with a real-world dataset covering a similar time span.

Several alternative methods were explored for the issues of predicting a variable multiple timesteps into the future and adapting a trained model to new data over time. The best overall performance for these issues was found with a recurrent NARX neural network that was retrained in a batch at the end of each day. One of the issues with this system was the tendency for errors to accumulate on successive predictions: this issue was addressed to some extent with the hybrid model method based on the calculated uncertainty level.

Investigation into the most effective input variables for the trained model showed a general tendency towards overfitting, as with the detection model. This meant that, in order to achieve a reasonable performance within a 3-week initial training period, the number of input variables and model structure needed to be kept to a minimal level. Time-based data most useful to pattern prediction was the time of day, with some limited value to the day of week or weekend/weekday indicator. Including the binary presence of occupants was found to be too much data for a system to train quickly – it caused overfitting over a full 100-day training period and so was discarded from the final model. The number of past timesteps included needed a balance between keeping the number of inputs low to avoid overfitting and including enough past context to avoid issues around expecting the correct number of occupants back after absences.

As it was possible to manipulate the complexity of occupancy patterns in the generated training data, it was investigated how the trained model performed against simpler prediction techniques as the data became more realistically complex. It was found that the model gave a better performance than the baseline fixed occupancy schedule up to 110 minutes into the future when trained on a realistic dataset.

The final proposed predictive model was a hybrid approach that avoided some of the more extreme errors from the trained neural network by deferring to a simpler mean-based prediction when the calculated uncertainty level was beyond a given threshold. This threshold was determined by the emphasis on improving the performance in unusual occupancy situations, but could be altered depending on the priorities required by the building energy management system in application to real building controls.

The system developed in this chapter is designed to facilitate the proactive control of slower-response systems such as heating and ventilation. The prediction horizon of interest was set at 2 hours in this work, but the most useful prediction horizon in application would depend on the characteristics of the space, the system to be controlled and the gap between its starting state and desired state. For many applications, a prediction horizon lower than two hours would be sufficient.

In application, a factor of interest is how failures of the predictive model could be handled, and what consequences could arise from a misprediction. An incorrect

prediction could be caused in two different ways: failure of the model itself, or incorrect inputs from the detection model.

Failure of the predictive model itself – as there will always be some unpredictability in real occupant patterns, even a predictive model that could perfectly represent the patterns in its training data would not be able to perfectly anticipate future events at all times, and so mispredictions should be expected within normal running conditions. The predicted uncertainty level should allow for informed treatment of many such situations, as higher errors were typically associated with higher uncertainty, as shown by the improved performance of the hybrid model in section 7.11. However, there will also be some situations where occupants diverge from their typical patterns and a higher-certainty prediction will prove to be false. In these cases, the control outcome will depend on the priorities of the building manager: control logic could be implemented to enact more aggressive energy saving measures, where a failed prediction would result in less comfortable occupant conditions, or if occupant comfort takes a higher priority, the space conditions could be kept closer to comfort levels at the expense of losing some smaller energy saving opportunities.

Failure of occupant data supplied to the predictive model – to allow for systematic analysis of the prediction model developed in this chapter, the assumption had to be made that the training data supplied was representative of actual occupancy rates. The work in Chapter 6 suggests that this will not always be the case: occupancy detection systems that count all occupants of a space were not always correct. If the prediction model is working on incorrect data, there is potential for a greater deviation from actual future occupancy rates. Once again, the Bayesian treatment of the predictive model is designed to counter these issues to some extent: the possibility of random noise on the prediction model inputs is accounted for in the uncertainty method used. However, in the case of systematic errors, which could feasibly be encountered where the detection data side has either been poorly trained or encounters sensor data far outside the scope of its original training, this would represent a failure for the predictive model where errors from the detection model could be compounded. This is one of the drawbacks to splitting the job of occupant detection and prediction into two separate modules. Where a predictive model is consistently producing poor outputs due to poor input data, a manual intervention may be required to replace or re-train the occupant detection source. As part of a greater control system, it may be

possible to include logic where manual interventions to the building controls could prompt an automated assessment and re-train of the occupancy sensing models to attempt to reduce such errors over time. This possibility is beyond the scope of this work, but would be a valuable consideration for further study.

8 DISCUSSION AND CONCLUSIONS

The body of this work covers the development of systems and methods to address uncertainties around building occupancy through the collection and processing of real-time occupant data for building energy management purposes. Through identifying areas of interest in case studies, systems targeted at a feasible and low-cost solution to occupant localisation were developed and tested in a small office setting. Models to combine sensor data into an estimation of local occupancy rates and to predict near-future changes in occupancy were developed through the application of machine learning techniques, with the aim of generating relevant and useable occupant information to better inform building automation systems and to target building energy use more closely towards actual occupant needs.

While there is a significant performance gap between designed and actual building energy use, in which occupant behavioural and building control issues are a significant factor, it was found that current building controls systems rarely explicitly measure occupant data, severely limiting their capacity to react appropriately to changing occupant needs and highlight potential places to save energy without negatively impacting occupant experience. In recent years, the Building Energy Management Systems (BEMS) field has seen an increasing interest in the intelligence included in building automation software.

From the wide range of approaches into the collection of occupant data undertaken in existing research, it can be seen that the measurement of building occupants is a highly complex problem that requires a clearly defined aim in the type and level of data collected, as well as a necessary trade-off between the level of detail measured and the perceived intrusion into occupants' privacy. 'Occupant data' could encompass a wide range of actual data types, from simple binary presence to highly computationally intensive systems to infer specific occupants' activities. In application to building controls, occupant data can be used in a range of ways, approximately falling into four categories: real-time response to occupant changes, control to individual preference, control based on activity type and pre-emptive predictive control. Due to the wide range of building types and control methods reviewed, it is difficult to directly compare the energy saving made by different studies. In research that compares different strategies in the same building, it appeared

that the greatest overall energy saving was achieved with controls that combine reactive and predictive approaches to optimise conditioning of a space, although some authors questioned the added value of prediction versus its increased computational requirements. The occupant-predictive energy management and control was identified as an area with a potential for expansion as available computing resources develop. This finding informed the emphasis on predictive algorithms for occupancy rate in the later stages of this work.

The overarching conclusion from the case studies presented in Chapter 4 is a confirmation of some of the issues around occupant data collection that were identified in the review chapters: the collection of any kind of in-use occupancy data is not widely implemented, and where data is collected, it is often not sufficient for a full understanding of where energy expenditure actually provides benefit to the building user. In a domestic setting, it was found that binary occupant presence can be inferred from single-point motion and CO_2 data, although with some reliability issues. Broad energy trends were analysed, finding that the heating and ventilation systems were often controlled without regard to the presence of the occupants. In a large office setting, access control data was used as a proxy for building occupancy, identifying some mismatch between the change of energy use and building population across weekdays and weekends. Different office floors showed different correlations to their daily occupancy rate, potentially highlighting more wasteful energy or control behaviours on these floors. In a school setting, the collection of explicit occupancy data ran into concerns about privacy. Through examination of predefined active school hours, it was found that occupant actions had a significant effect on the energy use relative to expected rates, with higher than expected occupancy outside of contracted hours raising the overall energy use above the accepted threshold for the building's energy contract. Through qualitative interviewing it was seen that the conditioning of some spaces did not adequately account for sudden large changes in occupancy, highlighting the need for slow-response building systems to be able to work pre-emptively where sudden changes in occupant rates can be anticipated.

The technologies set up and tested in Chapter 5 showed some successes in their goals of minimal-cost, passive collection of occupant data using wireless signals and users' personal devices. Using the single-point Raspberry Pi device to listen for local Wi-Fi probe signals, more generic information can be collected than with the more typical

method of counting the number of devices connected to a specific network, as this system also counts visitors who do not actively connect to any local Wi-Fi networks. Assumptions were applied to group together detection events for each device, allowing a continuous measure of presence. However, the frequency of detection was found to be highly variable depending on factors such as the type and use of the personal device. The signal strength of the received Wi-Fi signals was found to have a relation to locality, with a significant drop in signal strength seen caused by both distance and blocking by building fabric. This allows the signal strength to be used to filter out false positive results from people passing nearby the monitored area, but this is dependent on the location of the sensor and makeup of the building. The second system tested used software running on specific personal devices to detect the distance from a series of Bluetooth-LE beacons around the monitored building. An Android-based app was developed for this purpose. The use of Bluetooth-LE specifically allowed for more frequent probes with the same battery capacity, meaning that finer-grade data can be collected with less of an impact on the battery life of components. However, signal strength of Bluetooth was found to be highly variable due to physical properties of Bluetooth frequency waves: the signal was easily blocked by people, furniture and building fabric. The capture rate of both of the tested systems rely to some extent on properties of occupants' personal devices, which cannot be relied upon across an entire building population. This means that personal device detection in these forms would not currently be appropriate for use alone in a building energy management occupancy detection system, where reliability and consistency are important. However, as part of a greater system, these two measures can provide valuable information.

As a more comprehensive and uncertainty-sensitive system was needed, a system was proposed to combine sensor data from multiple sources, allowing benefit from the specific strengths of multiple sensor types. From the review of existing studies, it was decided that the most effective BEMS occupancy measure should also include some prediction of future events, remaining sensitive to the unavoidable uncertainties in future prediction. The task of interpreting this occupant data was split into two modules: the first focussing on occupancy detection, and the second focussing on probabilistic prediction.

The first stage of the data interpretation was the development of a model to interpret useable occupancy data from raw sensor data, which was detailed in Chapter 6. For this task, an Artificial Neural Network model was trained on labelled ground truth data, with the aim of combining raw sensor data to estimate a local number of occupants. It was found that zone-level networks trained on a selected subset of the available sensors provided the lowest error on the test data. Methods for pre-processing incoming data were tested, finding that the recent trend in local CO_2 was better able to represent occupancy rate than the absolute value in most zones. Noise removal on data inputs was found to be effective as long as the noise removal technique did not cause a significant time lag on changes in the data values. Data from personal mobile devices also required some pre-processing to ignore non-personal devices and to reduce the total number of inputs supplied to the model. The most effective combination of sensors varied from zone to zone, with reliance on the type and pattern of space use, as well as some physical properties of the zone. In application, it is essential that these differences are accounted for if a system or model has only been tested in one type of space. It was also found that some spaces did not see enough use during the training period to properly encode their patterns into the model. For more comprehensive research into the relationship between space use and the effectiveness of passive occupancy sensing methods, much broader datasets of diverse space types would be needed in testing.

A significant shortcoming found in the supervised learning approach used in this chapter was the reliance on manually labelled training data. As the data collection periods were short due to the practical issues around manual location recording, models had to be kept relatively simplistic to avoid the overfitting issues associated with a smaller training dataset. This issue is common to any supervised learning approach and cannot be avoided without introducing other potential sources of error, such as relying on the personal judgement that is typically needed to interpret useable information from unsupervised models.

Ultimately, it is expected that as technologies progress, an occupancy detection model based on the particular passive data sources tested in this work may become redundant. For example, if a more reliable system can be made using data from personal devices, this would perform mostly the same function at a considerably lower time and monetary cost than installing the full range of sensor types tested in

this work. However, it should be noted that some form of sensor combination is necessary if all occupant types are to be detected: not all occupants have mobile devices switched on at all times, not all occupants are regular office users that could be expected to carry tagging equipment. In long-term post-occupancy analysis or general pattern observation, there is less penalty if some occupants are missed. However, in building services control it is essential that the building can respond to any occupant. In theory, any reliable occupancy detection system could be used as a data supply for a predictive model for future occupancy, which is developed in the next chapter.

As the second stage of the occupant data interpretation, in Chapter 7 a predictive model was developed to pre-emptively estimate changes in local occupancy rate based on the recent past. A Bayesian Neural network was selected to allow for the calculation of uncertainty levels in the model output, with the acknowledgement that achieving a perfect prediction of future events is not possible, the level of certainty in the predictions made is a valuable measure for the purpose of making control decisions. The structure and optimisation of the proposed model was developed using a generated dataset to allow for testing on training data of an arbitrary length and control over the types of patterns shown in the data. A method to generate a dataset with same characteristics as the test building was developed as a part of this work, with a basis on using individual occupant agents with individual variations on characteristics such as typical times of occupancy and likelihood of staying longer than usual. A restriction was placed on the model performance that any model that required more than three weeks of training data to start giving reasonable results would be discarded even if it provided better results over longer periods, as it is unreasonable to expect a non-functional system to run in a building for months before it begins working as intended. This limited the complexity of patterns that the model was able to represent, but it was considered unrealistic to have a poor performance for too long in a live system. After optimisation and testing, it was found that the trained model gave a lower prediction error on the local number of occupants than the baseline fixed occupancy schedule up to 110 minutes into the future when trained on a realistic dataset. The final proposed predictive model was a hybrid approach that avoided some of the more extreme errors from the trained neural network by deferring to a simpler mean-based prediction when the calculated uncertainty level was beyond

a given threshold. This threshold was determined by the emphasis on improving the performance in unusual occupancy situations, but could be altered depending on the priorities required by the building energy management system in application to real building controls.

The predictive system developed was designed to facilitate the proactive control of slower-response systems such as heating and ventilation. The prediction horizon of interest was set at 2 hours in this work, but the most useful prediction horizon in application would depend on the characteristics of the space, the system to be controlled and the gap between its starting state and desired state.

Through each stage of this work, the findings have reinforced that while there is a clear need to increase the understanding around occupants and energy use and to increase the responsiveness of buildings to their occupants, the ways that occupant data can be collected are affected by a wide variety of application-specific challenges. The ways in which occupants can interact with built environments are as diverse as occupants themselves, and any system to tie this interaction down to a single measurable value will necessarily lose some nuance in doing so. It is therefore of the highest importance that occupancy sensing systems are designed sensitively to the desired application and the space being monitored. There are several, often competing, factors to consider:

- Application requirements – exactly what aspect of occupant data needs to be measured (presence, location, activity etc.) depends heavily on how the data is to be used. In controls, reliability of data collection is very important as there is a comfort or energy cost to making errors; for general trend analysis to feed back into designs, this is less so. For fast-response systems occupants must be detected instantly, but slow-response systems may need pre-emptive anticipation of patterns or prediction of future events.
- Privacy of occupants – both perceived and in terms of data security. Occupants may not be comfortable with systems perceived as intrusive, and where data is linked directly to a person's identity, participant consent must be sought and data must be appropriately protected. Passive data collection is not as likely to collect information that can be directly tied to one person, but this approach may not provide enough information for some controls, such as pre-

emptively switching on a user's office equipment as they are detected approaching.

- Installation costs and practicalities – in many cases, improved controls systems will be applied in existing buildings where it is not feasible to install a large number of wired sensors. Developments in wireless technology have allowed for a more viable approach to indoor data collection.
- Physical components – it was shown through this work that space characteristics can affect the way systems react to similar situations. In particular, the detection of personal devices through Wi-Fi and Bluetooth were found to be significantly affected by the building fabric and layout. Environmental sensors were also seen to have different responses depending on the characteristics of their location.
- Behavioural components – the response of physical sensors was also shown to be somewhat dependent on stay duration. Systems based on occupant participation depend heavily on the correct use of equipment from occupants – be that remembering to carry tags or ensuring that Bluetooth is enabled on their personal device.

The complicating factors for data measurement discussed above were part of the motivation to keep the data processing solutions developed in this work in a modular form: by keeping the sensor-end separated from the controls-facing predictive model, the system could theoretically allow for the introduction of new data sources or an updated way to interpret occupancy rates while maintaining the established learned occupancy patterns in the prediction model. This also highlights some benefits of building a system that makes use of multiple sensor types: where single sensors leave room for uncertainty, multiple sensors affected by different factors may be able to fill in gaps and create a more robust whole.

8.1 Recommendations for Future Work

The work presented in this thesis covers several stages of a greater process in the inclusion of occupant data into building energy management. The direct continuation of this work would be to apply the occupancy detection and prediction models developed here to a live building automation system: a process that would involve further levels of appropriate machine intelligence to make the most effective use of the local anticipated occupancy changes and their calculated uncertainty. While predictive controls are an area of increasing interest in the built environment, to the author's knowledge the application of explicitly calculated occupancy uncertainty would be a further novel step for building controls.

A major consideration for further study is the issue of how failures in the detection and prediction of occupancy rate could be appropriately handled by an energy management system. As mentioned in Chapter 7, the logic of how a control system responds to high uncertainty situations, as well as the priorities for whether unpredicted events end in energy wasted or deviation from comfort conditions depend on the specific requirements of an application and the priorities of the building manager.

In order to make the best use of localised occupant data, it is also essential that building controls are able to match the granularity of the observed occupant energy demand. A clear need has been established to tighten the use of building energy systems such as HVAC, lighting etc. to more closely meet what occupants actually need. This requires change on the control logic side, but also needs the physical capability of building controls to better localise building services. In many buildings services systems, large numbers of rooms can be fed by a single-actuator feed from a centralised control system. In this case, localised drops in energy requirement cannot be acted upon even if detected. To some extent, solutions to localised actuation of buildings services can be implemented on top of existing systems using controllable local actuators such as relays, local Wi-Fi-enabled sockets, local thermostats etc. For bigger systems, this change would need to be brought into the design stages of automation systems.

The systems tested as part of this work were assessed for their usefulness in control applications, but could also be applied to other post-occupancy evaluation studies or

real-time analytics. In particular, the Wi-Fi and Bluetooth-LE based personal device detection showed the ability to gather localised occupant counts and to follow the indoor location of specific occupants with relatively low-cost equipment. If applied more widely, these technologies could provide long-term occupant usage data to feed back into the design stages for new buildings, addressing one of the other contributing factors to the observed building performance gap.

8.2 Publications from this work

Some sections of the work in this thesis have been presented in published work. Further publications are pending and planned.

Current publications from this work:

- "A Post-Occupancy Case Study on the Relationship between Domestic Energy Use and Occupancy Profiles" [227] – presented at SET 2015 conference.
- "A concept review of power line communication in building energy management systems for the small to medium sized non-domestic built environment" [228] – Renewable and Sustainable Energy Reviews.
- "Bluetooth-based Mobile Device Detection for Improved Energy-Saving Controls System" [229] – presented at SET 2017 conference.
- "The Development of Occupancy Monitoring for Removing Uncertainty within Building Energy Management Systems" [230] – presented at ICL-GNSS 2017.

Pending publications:

- "A Review of Occupant-Centric Building Control Strategies to Reduce Building Energy Use" – Renewable and Sustainable Energy Reviews. Submission has received feedback from reviewers and amendments are in progress.

Planned Publications:

- "Developing a System for Localised Occupancy Detection for Building Energy Management" (Working Title).
- "Developing a System for Occupancy Prediction for Building Energy Management" (Working Title).

9 REFERENCES

[1] Department for Communities and Local Government, "Energy performance of buildings - Improving the energy efficiency of buildings and using planning to protect the environment," 05-Sep-2013. [Online]. Available: https://www.gov.uk/government/policies/improving-the-energy-efficiency-of-buildings-and-using-planning-to-protect-the-environment/supporting-pages/energy-performance-of-buildings. [Accessed: 03-Feb-2014].

[2] "Energy Consumption in the UK (2013) Chapter 1: Overall factsheet," Department of Energy & Climate Change, UK, 13D/154, Jul. 2013.

[3] E. Azar and C. C. Menassa, "A comprehensive analysis of the impact of occupancy parameters in energy simulation of office buildings," *Energy and Buildings*, vol. 55, pp. 841–853, Dec. 2012.

[4] CarbonBuzz, "Summary of Audits Performed on CarbonBuzz by the UCL Energy Institute," Apr-2013. [Online]. Available: http://www.carbonbuzz.org/downloads/PerformanceGap.pdf. [Accessed: 06-May-2014].

[5] M. Herrando, D. Cambra, M. Navarro, L. de la Cruz, G. Millán, and I. Zabalza, "Energy Performance Certification of Faculty Buildings in Spain: The gap between estimated and real energy consumption," *Energy Conversion and Management*.

[6] A. C. Menezes, A. Cripps, D. Bouchlaghem, and R. Buswell, "Predicted vs. actual energy performance of non-domestic buildings: Using post-occupancy evaluation data to reduce the performance gap," *Applied Energy*, vol. 97, pp. 355–364, Sep. 2012.

[7] R. Lowe and T. Oreszczyn, "Regulatory standards and barriers to improved performance for housing," *Energy Policy*, vol. 36, no. 12, pp. 4475–4481, Dec. 2008.

[8] CarbonBuzz, "CarbonBuzz - Benchmarks," 2014. [Online]. Available: http://www.carbonbuzz.org/index.jsp#performancegap. [Accessed: 06-May-2014].

[9] "Energy Consumption in the UK (2013) Chapter 3: Domestic Factsheet," Department of Energy & Climate Change, UK, 13D/158, Jul. 2013.

[10] I. Hamilton, P. Steadman, R. Lowe, A. Summerfield, and H. Bruhns, "Building and energy data frameworks: Report on the exploratory analysis of the Homes Energy Efficiency Database and Energy Demand," UCL Energy Institute, Oct. 2011.

[11] "English housing survey 2011 to 2012: headline report," Department for Communities and Local Government, ISBN 9781409837770, Jul. 2013.

[12] C. Randall, "Social Trends 41 - Housing," Office for National Statistics, ISSN 2040-1620, Feb. 2011.

[13] "Cavity wall insulation - Insulation - Homes - Energy Saving Trust." [Online]. Available: http://www.energysavingtrust.org.uk/Insulation/Cavity-wall-insulation. [Accessed: 20-Jan-2014].

[14] "How much does Home Automation really Cost?," *Home Automation Company UK*. [Online]. Available: http://www.homeautomationcompany.co.uk/blog/how-much-does-home-automation-cost/. [Accessed: 20-Jan-2014].

[15] "Energy Consumption in the UK (2013) Chapter 5: Services factsheet," Department of Energy & Climate Change, UK, 13D/162, Jul. 2013.

[16] "Energy Consumption in the UK (2013) Chapter 4: Industrial Factsheet," Department of Energy & Climate Change, UK, 13D/160, Jul. 2013.

[17] M. Wigginton and J. Harris, *Intelligent Skins*. Butterworth-Heinemann, 2002.

[18] J. K. W. Wong, H. Li, and S. W. Wang, "Intelligent building research: a review," *Automation in Construction*, vol. 14, no. 1, pp. 143–159, Jan. 2005.

[19] A. Mahdavi, Ed., "The technology of sentient buildings," 2006.

[20] "Stichting Smart Homes, Nationaal Kenniscentrum Domotica & Slim Wonen - Domotics." [Online]. Available: http://www.smart-homes.nl/domotica.aspx?lang=en-US. [Accessed: 20-Jan-2014].

[21] "What is smart home or building? - Definition from WhatIs.com," *IoT Agenda*. [Online]. Available: http://internetofthingsagenda.techtarget.com/definition/smart-home-or-building. [Accessed: 25-Mar-2018].

[22] S. Wang, *Intelligent Buildings and Building Automation*. Taylor & Francis, 2009.

[23] "Building Automation and Control System BACS - Designing Buildings Wiki." [Online]. Available: https://www.designingbuildings.co.uk/wiki/Building_Automation_and_Control_System_BACS. [Accessed: 25-Mar-2018].

[24] J. P. Fox and C. Wheelock, "Executive Summary: Building Energy Management Systems - Enabling Systems for Energy Efficiency, Demand Response, Energy Management, and Facility Automation in Commercial Buildings," Pike Research, 2010.

[25] D. Chakraborty, "Building Energy Management Systems," *Schneider Electric Blog*, 04-Apr-2017. [Online]. Available: https://blog.schneider-electric.com/building-management/2017/04/04/building-energy-management-systems/. [Accessed: 25-Mar-2018].

[26] "Module 8: Sensing the need for demand controlled ventilation," *CIBSE Journal*. .

[27] "Direct digital control (DDC) | Building System Design SolutionsBuilding System Design Solutions." [Online]. Available: http://www.bsdsolutions.com/about-us/bsd-news/tag/direct-digital-control-ddc/. [Accessed: 25-Mar-2018].

[28] J. Granderson, M. A. Piette, and G. Ghatikar, "Building energy information systems: user case studies," *Energy Efficiency*, vol. 4, no. 1, pp. 17–30, Jun. 2010.

[29] "Pros and Cons of Wireless Building Automation Systems - Facilities Management Building Automation Feature." [Online]. Available: http://www.facilitiesnet.com/buildingautomation/article/Pros-and-Cons-of-Wireless-Building-Automation-Systems--13856?source=part. [Accessed: 20-Jan-2014].

[30] P. Jones MEI, "Building Energy Management Systems (BEMS)," *Energy in Buildings and Industry CPD Scheme, Series 9, Module 05*, pp. 29–32, 2011.

[31] "SEAI - Building Energy Management Systems (BEMS)." [Online]. Available: http://www.seai.ie/Your_Business/Technology/Buildings/Building_Energy_Management_Systems_BEMS_.html. [Accessed: 20-Jan-2014].

[32] D. Bruckner and R. Velik, "Behavior Learning in Dwelling Environments With Hidden Markov Models," *IEEE Transactions on Industrial Electronics*, vol. 57, no. 11, pp. 3653–3660, Nov. 2010.

[33] P. H. Shaikh, N. B. M. Nor, P. Nallagownden, I. Elamvazuthi, and T. Ibrahim, "A review on optimized control systems for building energy and comfort management of smart sustainable buildings," *Renewable and Sustainable Energy Reviews*, vol. 34, pp. 409–429, Jun. 2014.

[34] Carbon Trust, “Building controls: Realising savings through the use of controls,” Technology Overview CTV032, 2007.
[35] “Oct. 29, 2013: Building Energy Management Systems Market to Reach $23 Billion by 2017.” [Online]. Available: http://www.achrnews.com/articles/124660-oct-29-2013-building-energy-management-systems-market-to-reach-23-billion-by-2017. [Accessed: 21-Jan-2014].
[36] “Honeywell Home,” *Honeywell Home*. [Online]. Available: http://lyric.honeywell.com/innovation/. [Accessed: 22-Jun-2015].
[37] “Honeywell Products & Services | Buildings, Construction, Maintenance.” [Online]. Available: http://honeywell.com/Products-Services/Pages/buildings-construction-maintenance.aspx. [Accessed: 21-Jan-2014].
[38] “Johnson Controls.” [Online]. Available: http://www.johnsoncontrols.com/. [Accessed: 25-Mar-2018].
[39] “Global Specialist in Energy Management and Automation | Schneider Electric.” [Online]. Available: https://www.schneider-electric.co.uk/en/. [Accessed: 25-Mar-2018].
[40] “Home - English - United Kingdom.” [Online]. Available: https://www.siemens.com/uk/en/home.html. [Accessed: 25-Mar-2018].
[41] “Home,” *United Technologies*. [Online]. Available: http://www.utc.com/Pages/Home.aspx. [Accessed: 25-Mar-2018].
[42] “Best Home Automation UK 2016 - A Detailed Comparison.” [Online]. Available: http://www.appcessories.co.uk/best-home-automation-uk-system/. [Accessed: 07-Mar-2016].
[43] “Hive Thermostat & Heating Control App | Hive Active Heating.” [Online]. Available: https://www.hivehome.com/hive-active-heating. [Accessed: 30-Oct-2015].
[44] “Meet the Nest Thermostat,” *Nest*. [Online]. Available: https://nest.com/uk/thermostat/meet-nest-thermostat/. [Accessed: 30-Oct-2015].
[45] “Heat Genius Products and Services: Heating App, Smart Thermostats.” [Online]. Available: https://www.heatgenius.co.uk/products/. [Accessed: 30-Oct-2015].
[46] “For Heating & Air Conditioning – Intelligent Climate Control,” *tado°*. [Online]. Available: https://www.tado.com/gb. [Accessed: 30-Oct-2015].
[47] “evohome - Honeywell UK Heating Controls.” [Online]. Available: http://www.honeywelluk.com/products/Underfloor-Heating/evohome-Main/. [Accessed: 30-Oct-2015].
[48] J. Martin, “Nest’s latest thermostat can control your hot water as well as your heating,” *Tech Advisor*, 10-Mar-2017. [Online]. Available: http://www.techadvisor.co.uk/review/smart-thermostats/nest-thermostat-review-3rd-gen-3543915/. [Accessed: 24-Sep-2017].
[49] C. M. Clevenger and J. Haymaker, “The impact of the building occupant on energy modeling simulations,” in *Joint International Conference on Computing and Decision Making in Civil and Building Engineering, Montreal, Canada*, 2006, pp. 1–10.
[50] M. Eguaras-Martínez, M. Vidaurre-Arbizu, and C. Martín-Gómez, “Simulation and evaluation of Building Information Modeling in a real pilot site,” *Applied Energy*, vol. 114, pp. 475–484, Feb. 2014.
[51] E. Azar and C. Menassa, “Agent-Based Modeling of Occupants and Their Impact on Energy Use in Commercial Buildings,” *Journal of Computing in Civil Engineering*, vol. 26, no. 4, pp. 506–518, 2012.
[52] C. Martani, D. Lee, P. Robinson, R. Britter, and C. Ratti, “ENERNET: Studying the dynamic relationship between building occupancy and energy consumption,” *Energy and Buildings*, vol. 47, pp. 584–591, Apr. 2012.
[53] O. T. Masoso and L. J. Grobler, “The dark side of occupants’ behaviour on building energy use,” *Energy and Buildings*, vol. 42, no. 2, pp. 173–177, Feb. 2010.
[54] Z. Yu, B. C. M. Fung, F. Haghighat, H. Yoshino, and E. Morofsky, “A systematic procedure to study the influence of occupant behavior on building energy consumption,” *Energy and Buildings*, vol. 43, no. 6, pp. 1409–1417, Jun. 2011.
[55] Y. G. Yohanis, J. D. Mondol, A. Wright, and B. Norton, “Real-life energy use in the UK: How occupancy and dwelling characteristics affect domestic electricity use,” *Energy and Buildings*, vol. 40, no. 6, pp. 1053–1059, Jan. 2008.
[56] T. S. Blight and D. A. Coley, “Sensitivity analysis of the effect of occupant behaviour on the energy consumption of passive house dwellings,” *Energy and Buildings*, vol. 66, pp. 183–192, Nov. 2013.
[57] H. B. Gunay, W. O’Brien, and I. Beausoleil-Morrison, “A critical review of observation studies, modeling, and simulation of adaptive occupant behaviors in offices,” *Building and Environment*, vol. 70, pp. 31–47, Dec. 2013.
[58] A. Mahdavi and C. Pröglhöf, “User behavior and energy performance in buildings,” *Wien, Austria: Internationalen Energiewirtschaftstagung an der TU Wien (IEWT)*, 2009.
[59] A. H. Kazmi, M. J. O’grady, D. T. Delaney, A. G. Ruzzelli, and G. M. P. O’hare, “A Review of Wireless-Sensor-Network-Enabled Building Energy Management Systems,” *ACM Trans. Sen. Netw.*, vol. 10, no. 4, pp. 66:1–66:43, Jun. 2014.
[60] T. Yu, “Modeling Occupancy Behavior for Energy Efficiency and Occupants Comfort Management in Intelligent Buildings,” 2010, pp. 726–731.
[61] S. N. Patel, M. S. Reynolds, and G. D. Abowd, “Detecting Human Movement by Differential Air Pressure Sensing in HVAC System Ductwork: An Exploration in Infrastructure Mediated Sensing,” in *Pervasive Computing*, J. Indulska, D. J. Patterson, T. Rodden, and M. Ott, Eds. Springer Berlin Heidelberg, 2008, pp. 1–18.
[62] E. Hailemariam, R. Goldstein, R. Attar, and A. Khan, “Real-time occupancy detection using decision trees with multiple sensor types,” in *Proceedings of the 2011 Symposium on Simulation for Architecture and Urban Design*, 2011, pp. 141–148.
[63] L. M. Candanedo, V. Feldheim, and D. Deramaix, “A methodology based on Hidden Markov Models for occupancy detection and a case study in a low energy residential building,” *Energy and Buildings*, vol. 148, pp. 327–341, Aug. 2017.
[64] D. Chen, S. Barker, A. Subbaswamy, D. Irwin, and P. Shenoy, “Non-Intrusive Occupancy Monitoring using Smart Meters,” 2013, pp. 1–8.
[65] E. Naghiyev, M. Gillott, and R. Wilson, “Three unobtrusive domestic occupancy measurement technologies under qualitative review,” *Energy and Buildings*, vol. 69, pp. 507–514, Feb. 2014.
[66] M. Georgescu and I. Mezic, “Estimating Occupancy States from Building Temperature Data using Wavelet Analysis,” in *BS2013*, Le Bourget Du Lac, France, 2013.
[67] T. Ekwevugbe, N. Brown, and D. Fan, “Using indoor climatic measurements for occupancy monitoring,” 2012.
[68] S. Meyn, A. Surana, Y. Lin, S. M. Oggianu, S. Narayanan, and T. A. Frewen, “A sensor-utility-network method for estimation of occupancy distribution in buildings,” in *Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on*, 2009, pp. 1494–1500.
[69] O. İçoğlu and A. Mahdavi, “VIOLAS: A vision-based sensing system for sentient building models,” *Automation in Construction*, vol. 16, no. 5, pp. 685–712, Aug. 2007.

[70] A. Ebadat, G. Bottegal, D. Varagnolo, B. Wahlberg, and K. H. Johansson, “Estimation of building occupancy levels through environmental signals deconvolution,” 2013, pp. 1–8.
[71] B. Dong *et al.*, “An information technology enabled sustainability test-bed (ITEST) for occupancy detection through an environmental sensing network,” *Energy and Buildings*, vol. 42, no. 7, pp. 1038–1046, Jul. 2010.
[72] K. P. Lam *et al.*, “Information-theoretic environmental features selection for occupancy detection in open offices,” in *Eleventh International IBPSA Conference, edited by PA Strachan, NJ Kelly and M Kummert*, 2009, pp. 1460–1467.
[73] Z. Yang, N. Li, B. Becerik-Gerber, and M. Orosz, “A multi-sensor based occupancy estimation model for supporting demand driven HVAC operations,” in *Proceedings of the 2012 Symposium on Simulation for Architecture and Urban Design*, 2012, p. 2.
[74] T. Ekwevugbe, N. Brown, V. Pakka, and D. Fan, “Improved occupancy monitoring in non-domestic buildings,” *Sustainable Cities and Society*, vol. 30, pp. 97–107, Apr. 2017.
[75] K. Curran, E. Furey, T. Lunney, J. Santos, D. Woods, and A. Mc Caughey, “An Evaluation of Indoor Location Determination Technologies,” *Journal of Location Based Services*, vol. 5, no. 2, pp. 61–78, Jun. 2011.
[76] A. Misra and S. K. Das, “Location Estimation (Determination and Prediction) Techniques in Smart Environments,” in *Smart Environments : Technology, Protocols and Applications*, Hoboken, NJ, USA: John Wiley & Sons, Incorporated, 2004, pp. 193–228.
[77] M. Gillott, R. Holland, S. Riffat, and J. A. Fitchett, “Post-occupancy evaluation of space use in a dwelling using RFID tracking,” *Architectural Engineering and Design Management*, vol. 2, no. 4, pp. 273–288, 2006.
[78] N. Li, G. Calis, and B. Becerik-Gerber, “Measuring and monitoring occupancy with an RFID based system for demand-driven HVAC operations,” *Automation in Construction*, vol. 24, pp. 89–99, Jul. 2012.
[79] Z.-N. Zhen, Q.-S. Jia, C. Song, and X. Guan, “An indoor localization algorithm for lighting control using RFID,” in *Energy 2030 Conference, 2008. ENERGY 2008. IEEE*, 2008, pp. 1–6.
[80] C. Spataru and M. Gillott, “The use of intelligent systems for monitoring energy use and occupancy in existing homes,” in *Sustainability in Energy and Buildings*, Springer, 2011, pp. 247–256.
[81] C. Spataru, M. Gillott, and M. R. Hall, “Domestic energy and occupancy: a novel post-occupancy evaluation study,” *International Journal of Low-Carbon Technologies*, vol. 5, no. 3, pp. 148–157, Jul. 2010.
[82] R. Shipman and M. Gillott, “A Study of the Use of Wireless Behavior Systems to Encourage Energy Efficiency in Domestic Properties,” presented at the 12th International Conference on Sustainable Energy technologies (SET-2013), Hong Kong, 2013.
[83] R. H. Dodier, G. P. Henze, D. K. Tiller, and X. Guo, “Building occupancy detection through sensor belief networks,” *Energy and Buildings*, vol. 38, no. 9, pp. 1033–1043, Sep. 2006.
[84] S. Hay and R. Harle, “Bluetooth tracking without discoverability,” in *Location and context awareness*, Springer, 2009, pp. 120–137.
[85] Y. Zhao, W. Zeiler, G. Boxem, and T. Labeodan, “Virtual occupancy sensors for real-time occupancy information in buildings,” *Building and Environment*, vol. 93, pp. 9–20, Nov. 2015.
[86] J. Jun, L. Cheng, J. Sun, Y. Gu, T. Zhu, and T. He, “Improving Indoor Localization with Social Interactions,” in *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems*, New York, NY, USA, 2012, pp. 323–324.
[87] D. Lee and S. Oh, “Understanding human-place interaction from tracking and identification of many users,” in *Cyber-Physical Systems, Networks, and Applications (CPSNA), 2013 IEEE 1st International Conference on*, 2013, pp. 112–115.
[88] Z. B. Zhao, W. S. Xu, and D. Z. Cheng, “User behavior detection framework based on NBP for energy efficiency,” *Automation in Construction*, vol. 26, pp. 69–76, Oct. 2012.
[89] T. A. Nguyen and M. Aiello, “Beyond Indoor Presence Monitoring with Simple Sensors.,” in *PECCS*, 2012, pp. 5–14.
[90] F. J. Fernandez-Luque, F. Martínez, G. Domènech, J. Zapata, and R. Ruiz, “EMFi-based low-power occupancy sensor,” *Sensors and Actuators A: Physical*, vol. 191, pp. 78–88, Mar. 2013.
[91] F. Sebbak, A. Chibani, Y. Amirat, A. Mokhtari, and F. Benhammadi, “An evidential fusion approach for activity recognition in ambient intelligence environments,” *Robotics and Autonomous Systems*, vol. 61, no. 11, pp. 1235–1245, Nov. 2013.
[92] J. L. G. Ortega, L. Han, N. Whittacker, and N. Bowring, “A machine-learning based approach to model user occupancy and activity patterns for energy saving in buildings,” in *Science and Information Conference (SAI), 2015*, 2015, pp. 474–482.
[93] L. I. L. Gonzalez, M. Troost, and O. Amft, “Using a Thermopile Matrix Sensor to Recognize Energy-related Activities in Offices,” *Procedia Computer Science*, vol. 19, pp. 678–685, Jan. 2013.
[94] M. Philipose *et al.*, “Inferring activities from interactions with objects,” *IEEE Pervasive Computing*, vol. 3, no. 4, pp. 50–57, Oct. 2004.
[95] Y. Benezeth, H. Laurent, B. Emile, and C. Rosenberger, “Towards a sensor for detecting human presence and characterizing activity,” *Energy and Buildings*, vol. 43, no. 2–3, pp. 305–314, Feb. 2011.
[96] S. Mahmoud, A. Lotfi, and C. Langensiepen, “Behavioural pattern identification and prediction in intelligent environments,” *Applied Soft Computing*, vol. 13, no. 4, pp. 1813–1822, Apr. 2013.
[97] V. M. Zavala, “Proactive energy management for high-performance buildings: Exploiting and motivating sensor technologies,” in *Future of Instrumentation International Workshop (FIIW), 2011*, 2011, pp. 12–15.
[98] T. A. Nguyen and M. Aiello, “Energy intelligent buildings based on user activity: A survey,” *Energy and Buildings*, vol. 56, pp. 244–257, Jan. 2013.
[99] V. Garg and N. K. Bansal, “Smart occupancy sensors to reduce energy consumption,” *Energy and Buildings*, vol. 32, no. 1, pp. 81–87, 2000.
[100] T. Labeodan, C. De Bakker, A. Rosemann, and W. Zeiler, “On the application of wireless sensors and actuators network in existing buildings for occupancy detection and occupancy-driven lighting control,” *Energy and Buildings*, vol. 127, pp. 75–83, Sep. 2016.
[101] Y. Xu, N. Stojanovic, L. Stojanovic, D. Anicic, and R. Studer, “An approach for more efficient energy consumption based on real-time situational awareness,” in *The Semanic Web: Research and Applications*, Springer, 2011, pp. 270–284.
[102] D. J. Cook *et al.*, “MavHome: an agent-based smart home,” in *Proceedings of the First IEEE International Conference on Pervasive Computing and Communications, 2003. (PerCom 2003)*, 2003, pp. 521–524.
[103] S. Park, M. Choi, B. Kang, and S. Park, “Design and Implementation of Smart Energy Management System for Reducing Power Consumption Using ZigBee Wireless Communication Module,” *Procedia Computer Science*, vol. 19, pp. 662–668, Jan. 2013.

[104] C. Harris and V. Cahill, “Exploiting user behaviour for context-aware power management,” in *Wireless And Mobile Computing, Networking And Communications, 2005.(WiMob’2005), IEEE International Conference on*, 2005, vol. 4, pp. 122–130.
[105] N. Batra, P. Arjunan, A. Singh, and P. Singh, “Experiences with Occupancy Based Building Management Systems,” 2013.
[106] S. Meyer and A. Rakotonirainy, “A survey of research on context-aware homes,” in *Proceedings of the Australasian information security workshop conference on ACSW frontiers 2003-Volume 21*, 2003, pp. 159–168.
[107] Z. Sun, S. Wang, and Z. Ma, “In-situ implementation and validation of a CO2-based adaptive demand-controlled ventilation strategy in a multi-zone office building,” *Building and Environment*, vol. 46, no. 1, pp. 124–133, Jan. 2011.
[108] Z. Yang, N. Li, B. Becerik-Gerber, and M. Orosz, “A systematic approach to occupancy modeling in ambient sensor–rich buildings,” *Simulation*, vol. 90, no. 8, pp. 960–977, Jul. 2013.
[109] Z. Yang and B. Becerik-Gerber, “Modeling Personalized Occupancy Profiles for Representing Long Term Patterns by Using Ambient Context,” *Building and Environment*, 2014.
[110] Y. Agarwal, B. Balaji, R. Gupta, J. Lyles, M. Wei, and T. Weng, “Occupancy-driven energy management for smart building automation,” in *Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building*, 2010, pp. 1–6.
[111] M. Gruber, A. Trüschel, and J.-O. Dalenbäck, “Alternative strategies for supply air temperature control in office buildings,” *Energy and Buildings*.
[112] Y. Agarwal, B. Balaji, S. Dutta, R. K. Gupta, and T. Weng, “Duty-cycling buildings aggressively: The next frontier in HVAC control,” in *2011 10th International Conference on Information Processing in Sensor Networks (IPSN)*, 2011, pp. 246–257.
[113] B. Balaji, J. Xu, A. Nwokafor, R. Gupta, and Y. Agarwal, “Sentinel: occupancy based HVAC actuation using existing WiFi infrastructure within commercial buildings,” in *Proceedings of the 11th ACM Conference on Embedded Netwroked Sensor Systems*, 2013.
[114] W. Zeiler, G. Boxem, and R. Maaijen, “Wireless Sensor Technology to Optimize the Occupant’s Dynamic Demand Pattern within the Building,” 2012.
[115] S. Rosiek and F. J. Batlles, “Reducing a solar-assisted air-conditioning system’s energy consumption by applying real-time occupancy sensors and chilled water storage tanks throughout the summer: A case study,” *Energy Conversion and Management*, vol. 76, pp. 1029–1042, Dec. 2013.
[116] V. Singhvi, A. Krause, C. Guestrin, J. H. Garrett Jr, and H. S. Matthews, “Intelligent light control using sensor networks,” in *Proceedings of the 3rd international conference on Embedded networked sensor systems*, 2005, pp. 218–229.
[117] R. de Dear and G. S. Brager, “Developing an adaptive model of thermal comfort and preference,” Jan. 1998.
[118] J. Zhao, K. P. Lam, B. E. Ydstie, and V. Loftness, “Occupant-oriented mixed-mode EnergyPlus predictive control simulation,” *Energy and Buildings*.
[119] D. Kolokotsa, G. Saridakis, A. Pouliezos, and G. S. Stavrakakis, “Design and installation of an advanced EIB ™ fuzzy indoor comfort controller using Matlab ™,” *Energy and Buildings*, vol. 38, no. 9, pp. 1084–1092, 2006.
[120] P. X. Gao and S. Keshav, “SPOT: a smart personalized office thermal control system,” in *Proceedings of the fourth international conference on Future energy systems*, 2013, pp. 237–246.
[121] F. Jazizadeh, A. Ghahramani, B. Becerik-Gerber, T. Kichkaylo, and M. Orosz, “User-led decentralized thermal comfort driven HVAC operations for improved efficiency in office buildings,” *Energy and Buildings*, vol. 70, pp. 398–410, Feb. 2014.
[122] C. Y. Yong *et al.*, “Co-ordinated management of intelligent pervasive spaces,” in *Industrial Informatics, 2007 5th IEEE International Conference on*, 2007, vol. 1, pp. 529–534.
[123] H. Hagras, V. Callaghan, M. Colley, G. Clarke, A. Pounds-Cornish, and H. Duman, “Creating an ambient-intelligence environment using embedded agents,” *Intelligent Systems, IEEE*, vol. 19, no. 6, pp. 12–20, 2004.
[124] H. Chen, P. Chou, S. Duri, H. Lei, and J. Reason, “The Design and Implementation of a Smart Building Control System,” 2009, pp. 255–262.
[125] M. V. Moreno, M. A. Zamora, and A. F. Skarmeta, “User-centric smart buildings for energy sustainable smart cities,” *Transactions on Emerging Telecommunications Technologies*, p. n/a–n/a, 2013.
[126] M. V. Moreno-Cano, J. Santa, M. A. Zamora, and A. F. S. Gómez, “Context-Aware Energy Efficiency in Smart Buildings,” in *Ubiquitous Computing and Ambient Intelligence. Context-Awareness and Context-Driven Interaction*, Springer, 2013, pp. 1–8.
[127] M. V. Moreno-Cano, M. A. Zamora-Izquierdo, J. Santa, and A. F. Skarmeta, “An Indoor Localization System Based on Artificial Neural Networks and Particle Filters Applied to Intelligent Buildings,” *Neurocomput.*, vol. 122, pp. 116–125, Dec. 2013.
[128] L.-W. Yeh, Y.-C. Wang, and Y.-C. Tseng, “iPower: an energy conservation system for intelligent buildings by wireless sensor networks,” *International Journal of Sensor Networks*, vol. 5, no. 1, pp. 1–10, Jan. 2009.
[129] D. Vissers and W. Zeiler, “The User as Sensor to Reach for Optimal Individual Comfort and Reduced Energy Consumption,” in *Opportunities, Limits & Needs Towards an environmentally responsible architecture*, Lima , Perú, 2012.
[130] H. Lee, C. Wu, and H. Aghajan, “Vision-based user-centric light control for smart environments,” *Pervasive and Mobile Computing*, vol. 7, no. 2, pp. 223–240, Apr. 2011.
[131] M. Milenkovic and O. Amft, “An opportunistic activity-sensing approach to save energy in office buildings,” in *Proceedings of the fourth international conference on Future energy systems*, 2013, pp. 247–258.
[132] V. Pallotta, P. Bruegger, and B. Hirsbrunner, “Smart heating systems: Optimizing heating systems by kinetic-awareness,” in *Third International Conference on Digital Information Management, 2008. ICDIM 2008*, 2008, pp. 887–892.
[133] R. K. Harle and A. Hopper, “The potential for location-aware power management,” in *Proceedings of the 10th international conference on Ubiquitous computing*, 2008, pp. 302–311.
[134] “The adaptive house.” [Online]. Available: http://www.cs.colorado.edu/~mozer/index.php?dir=/Research/Projects/Adaptive%20house/. [Accessed: 14-Jan-2014].
[135] M. C. Mozer, “Lessons from an adaptive home,” in *Smart Environments : Technology, Protocols and Applications*, Hoboken, NJ, USA: John Wiley & Sons, Incorporated, 2004, pp. 273–298.
[136] D. J. Cook, G. M. Youngblood, and G. Jain, “Algorithms for smart spaces,” *The Engineering Handbook of Smart Technology for Aging, Dissablity and Independence, Wiley*, pp. 783–800, 2008.

[137] S. Mamidi, Y.-H. Chang, and R. Maheswaran, “Smart Sensing, Estimation, and Prediction for Efficient Building Energy Management,” in *Multi-agent Smart Computing Workshop*, 2011.
[138] S. Mamidi, Y.-H. Chang, and R. Maheswaran, “Improving Building Energy Efficiency with a Network of Sensing, Learning and Prediction Agents,” in *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*, Richland, SC, 2012, pp. 45–52.
[139] J. Howard and W. Hoff, “Forecasting building occupancy using sensor network data,” in *Proceedings of the 2nd International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*, 2013, pp. 87–94.
[140] Y. Peng, A. Rysanek, Z. Nagy, and A. Schlüter, “Occupancy learning-based demand-driven cooling control for office spaces,” *Building and Environment*, vol. 122, pp. 145–160, Sep. 2017.
[141] B. Dong and B. Andrews, “Sensor-based occupancy behavioral pattern recognition for energy and comfort management in intelligent buildings,” in *Proc. Int. IBPSA Conf*, 2009.
[142] B. Dong, K. P. Lam, and C. Neuman, “Integrated building control based on occupant behavior pattern detection and local weather forecasting,” in *Twelfth International IBPSA Conference. Sydney: IBPSA Australia*, 2011, pp. 14–17.
[143] B. Dong and K. P. Lam, “A real-time model predictive control for building heating and cooling systems based on the occupancy behavior pattern detection and local weather forecasting,” *Building Simulation*, vol. 7, no. 1, pp. 89–106, Sep. 2013.
[144] H. Burak Gunay, W. O’Brien, and I. Beausoleil-Morrison, “Development of an occupancy learning algorithm for terminal heating and cooling units,” *Building and Environment*, vol. 93, Part 2, pp. 71–85, Nov. 2015.
[145] J. R. Dobbs and B. M. Hencey, “Model Predictive HVAC Control with Online Occupancy Model,” *arXiv:1403.4662 [cs]*, Mar. 2014.
[146] A. Aswani, N. Master, J. Taneja, D. Culler, and C. Tomlin, “Reducing Transient and Steady State Electricity Consumption in HVAC Using Learning-Based Model-Predictive Control,” *Proceedings of the IEEE*, vol. 100, no. 1, pp. 240–253, Jan. 2012.
[147] A. Aswani, N. Master, J. Taneja, A. Krioukov, D. Culler, and C. Tomlin, “Energy-efficient building hvac control using hybrid system lbmpc,” *arXiv preprint arXiv:1204.4717*, 2012.
[148] A. Barbato, L. Borsani, A. Capone, and S. Melzi, “Home energy saving through a user profiling system based on wireless sensors,” in *Proceedings of the First ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, 2009, pp. 49–54.
[149] S. Lee, Y. Chon, Y. Kim, R. Ha, and H. Cha, “Occupancy prediction algorithms for thermostat control systems using mobile devices,” *Smart Grid, IEEE Transactions on*, vol. 4, no. 3, pp. 1332–1340, 2013.
[150] V. L. Erickson *et al.*, “Energy efficient building environment control strategies using real-time occupancy measurements,” in *Proceedings of the First ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, 2009, pp. 19–24.
[151] A. Kamthe, V. Erickson, M. A. Carreira-Perpinán, and A. Cerpa, “Enabling building energy auditing using adapted occupancy models,” in *Proceedings of the Third ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, 2011, pp. 31–36.
[152] V. L. Erickson and A. E. Cerpa, “Occupancy based demand response HVAC control strategy,” in *Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building*, 2010, pp. 7–12.
[153] V. L. Erickson, M. A. Carreira-Perpinan, and A. E. Cerpa, “OBSERVE: Occupancy-based system for efficient reduction of HVAC energy,” in *Information Processing in Sensor Networks (IPSN), 2011 10th International Conference on*, 2011, pp. 258–269.
[154] V. L. Erickson, S. Achleitner, and A. E. Cerpa, “POEM: power-efficient occupancy-based energy management system,” in *Proceedings of the 12th international conference on Information processing in sensor networks*, 2013, pp. 203–216.
[155] X. Zhang, G. Schildbach, D. Sturzenegger, and M. Morari, “Scenario-based MPC for energy-efficient building climate control under weather and occupancy uncertainty,” in *Control Conference (ECC), 2013 European*, 2013, pp. 1029–1034.
[156] F. Oldewurtel, D. Sturzenegger, and M. Morari, “Importance of occupancy information for building climate control,” *Applied Energy*, vol. 101, pp. 521–532, Jan. 2013.
[157] D. Sturzenegger, F. Oldewurtel, and M. Morari, “Importance of Long-Term Occupancy Information–A Validation with Real Occupancy Data,” in *Clima-RHEVA World Congress*, 2013.
[158] S. Goyal, H. A. Ingley, and P. Barooah, “Zone-level control algorithms based on occupancy information for energy efficient buildings,” in *American Control Conference (ACC), 2012*, 2012, pp. 3063–3068.
[159] S. Goyal, H. A. Ingley, and P. Barooah, “Occupancy-based zone-climate control for energy-efficient buildings: Complexity vs. performance,” *Applied Energy*, vol. 106, pp. 209–221, Jun. 2013.
[160] S. Goyal, P. Barooah, and T. Middelkoop, “Experimental study of occupancy-based control of HVAC zones,” 2014.
[161] S. Goyal, H. A. Ingley, and P. Barooah, “Effect of various uncertainties on the performance of occupancy-based optimal control of HVAC zones,” in *2012 IEEE 51st Annual Conference on Decision and Control (CDC)*, 2012, pp. 7565–7570.
[162] M. Gruber, A. Trüschel, and J.-O. Dalenbäck, “Model-based controllers for indoor climate control in office buildings – Complexity and performance evaluation,” *Energy and Buildings*, vol. 68, pp. 213–222, Jan. 2014.
[163] J. Lu *et al.*, “The smart thermostat: using occupancy sensors to save energy in homes,” in *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*, 2010, pp. 211–224.
[164] J. Widén and E. Wäckelgård, “A high-resolution stochastic model of domestic activity patterns and electricity demand,” *Applied Energy*, vol. 87, no. 6, pp. 1880–1892, Jun. 2010.
[165] R. Subbiah, K. Lum, A. Marathe, and M. Marathe, “Activity based energy demand modeling for residential buildings,” in *Innovative Smart Grid Technologies (ISGT), 2013 IEEE PES*, 2013, pp. 1–6.
[166] M. Muratori, M. C. Roberts, R. Sioshansi, V. Marano, and G. Rizzoni, “A highly resolved modeling technique to simulate residential power demand,” *Applied Energy*, vol. 107, pp. 465–473, Jul. 2013.
[167] D. Aerts, J. Minnen, I. Glorieux, I. Wouters, and F. Descamps, “A method for the identification and modelling of realistic domestic occupancy sequences for building energy demand simulations and peer comparison,” *Building and Environment*, vol. 75, pp. 67–78, May 2014.
[168] I. Richardson, M. Thomson, and D. Infield, “A high-resolution domestic building occupancy model for energy demand simulations,” *Energy and Buildings*, vol. 40, no. 8, pp. 1560–1566, Jan. 2008.
[169] I. Richardson, M. Thomson, D. Infield, and A. Delahunty, “Domestic lighting: A high-resolution energy demand model,” *Energy and Buildings*, vol. 41, no. 7, pp. 781–789, Jul. 2009.
[170] I. Richardson, M. Thomson, D. Infield, and C. Clifford, “Domestic electricity use: A high-resolution energy demand model,” *Energy and Buildings*, vol. 42, no. 10, pp. 1878–1887, Oct. 2010.

[171] H. Meidani and R. Ghanem, “Multiscale Markov models with random transitions for energy demand management,” *Energy and Buildings*, vol. 61, pp. 267–274, Jun. 2013.
[172] H. Meidani and R. Ghanem, “Uncertainty quantification for Markov chain models,” *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 22, no. 4, p. 043102, Oct. 2012.
[173] A. C. Menezes, A. Cripps, R. A. Buswell, J. Wright, and D. Bouchlaghem, “Estimating the energy consumption and power demand of small power equipment in office buildings,” *Energy and Buildings*, vol. 75, pp. 199–209, Jun. 2014.
[174] H. B. Rijal, P. Tuohy, M. A. Humphreys, J. F. Nicol, and A. Samuel, “An algorithm to represent occupant use of windows and fans including situation-specific motivations and constraints,” *Build. Simul.*, vol. 4, no. 2, pp. 117–134, Jun. 2011.
[175] W.-K. Chang and T. Hong, “Statistical analysis and modeling of occupancy patterns in open-plan offices using measured lighting-switch data,” *Build. Simul.*, vol. 6, no. 1, pp. 23–32, Mar. 2013.
[176] Y. S. Lee and A. Malkawi, “Simulating human behavior: an agent-based modeling approach,” in *Proceedings of the 13th IBPSA Conference, Chambéry*, 2013.
[177] Y. S. Lee and A. M. Malkawi, “Simulating multiple occupant behaviors in buildings: An agent-based modeling approach,” *Energy and Buildings*, vol. 69, pp. 407–416, Feb. 2014.
[178] C. Liao and P. Barooah, “A novel stochastic agent-based model of building occupancy,” in *American Control Conference (ACC), 2011*, 2011, pp. 2095–2100.
[179] C. Liao, Y. Lin, and P. Barooah, “Agent-based and graphical modelling of building occupancy,” *Journal of Building Performance Simulation*, vol. 5, no. 1, pp. 5–25, 2012.
[180] H. Brohus, C. Frier, P. Heiselberg, and F. Haghighat, “Quantification of uncertainty in predicting building energy consumption: A stochastic approach,” *Energy and Buildings*, vol. 55, pp. 127–140, Dec. 2012.
[181] J. Yang, H. Rivard, and R. Zmeureanu, “On-line building energy prediction using adaptive artificial neural networks,” *Energy and Buildings*, vol. 37, no. 12, pp. 1250–1259, Dec. 2005.
[182] “Natural Gas.” [Online]. Available: http://www.natural-gas.com.au/about/references.html. [Accessed: 17-Sep-2014].
[183] N. Ebbs, “Environmental Design in Architecture: Delivery of Sustainable Development -- Challenges and Opportunities. Green Street in the Meadows and Nottingham Waterside (Trent Basin),” 13-Oct-2011.
[184] S. Hone-Brookes and G. Mant, “Energy Platform Review - Dagenham Park Church of England School.” Laing O’Rourke Engineering Excellence Group, Nov-2013.
[185] “EnOcean CO2 Sensor Temperature and Humidity | Pressac,” *Pressac Communications*. [Online]. Available: http://www.pressac.com/enocean-co2-sensor-temperature-and-humidity. [Accessed: 25-Jan-2017].
[186] Niko-Servodan, “Tilstedeværelsessensor, solcelledrevet | Niko-Servodan.” [Online]. Available: http://www.niko.dk/side/tilstedevaerelsessensor-solcelledrevet. [Accessed: 25-Jan-2017].
[187] Thermokon Sensortechnik, “SR-MDS Solar - Light & Motion - Thermokon Sensortechnik.” [Online]. Available: http://www.thermokon.de/en/products/easysens-transmitters/light-motion/sr-mds-solar.html. [Accessed: 25-Jan-2017].
[188] Raspberry Pi Foundation, “Raspberry Pi 1 Model B+,” *Raspberry Pi*. .
[189] Kontakt, “Beacon,” 2016. [Online]. Available: https://store.kontakt.io/our-products/27-beacon.html. [Accessed: 25-Jan-2017].
[190] Kamstrup, “MULTICAL 402.” [Online]. Available: /en-eu/products-solutions/thermal-energy-meters/compact-meter-heat-and-cooling/multical-402. [Accessed: 25-Jan-2017].
[191] “Smartphone Tracking: How Close Is Too Close?” [Online]. Available: http://www.ecommercetimes.com/story/80251.html. [Accessed: 29-Jun-2015].
[192] “Compatibility Drivers Aircrack-ng.” [Online]. Available: https://www.aircrack-ng.org/doku.php?id=compatibility_drivers#which_is_the_best_card_to_buy. [Accessed: 08-Feb-2017].
[193] “Kismet.” [Online]. Available: http://www.kismetwireless.net/. [Accessed: 22-Jun-2015].
[194] M. Macleod, “Passive wifi presence detection using Raspberry Pi.” [Online]. Available: http://umm.io/blog/passive-wifi-tracking.html. [Accessed: 08-Feb-2017].
[195] V. Petrov, “BlueSense Beacon USB,” *Blue Sense Networks*, 14-Mar-2014. [Online]. Available: http://bluesensenetworks.com/product/bluebar-beacon-usb/. [Accessed: 15-Feb-2017].
[196] C. Hocking, “The Beacon Experiments: Low-Energy Bluetooth Devices in Action - DZone IoT,” *dzone.com*, Mar-2014. [Online]. Available: https://dzone.com/articles/beacon-experiments-low-energy. [Accessed: 15-Feb-2017].
[197] Aislelabs, “The Hitchhikers Guide to iBeacon Hardware: A Comprehensive Report by Aislelabs,” *Aislelabs*, 2015. .
[198] Tion, “Geofency for iPhone - Mobile Time Tracking,” 2013. [Online]. Available: http://www.geofency.com. [Accessed: 15-Feb-2017].
[199] Android Studio, “Android Studio and SDK Tools.” [Online]. Available: https://developer.android.com/studio/index.html. [Accessed: 20-Sep-2017].
[200] Radius Networks, “AltBeacon/android-beacon-library,” *GitHub*. [Online]. Available: https://github.com/AltBeacon/android-beacon-library. [Accessed: 27-Oct-2016].
[201] I. Bershadskiy, “How We Built iBeacon-based SilentBeacon App.” [Online]. Available: http://yalantis.com/blog/how-we-built-ibeacon-based-app-which-does-not-target-retail/. [Accessed: 10-Mar-2017].
[202] Beaconate.io, *Beacon Scanner & Logger (free)*. beaconate.io, 2015.
[203] Radius Networks, “Android Beacon Library,” 2014. [Online]. Available: http://altbeacon.github.io/android-beacon-library/distance-calculations.html. [Accessed: 27-Oct-2016].
[204] DigiKey, “Bluetooth 4.1, 4.2 and 5 SoCs Meet IoT Challenges,” 06-Apr-2017. [Online]. Available: https://www.digikey.co.uk/en/articles/techzone/2017/apr/bluetooth-41-42-5-low-energy-socs-meet-iot-challenges-part-1. [Accessed: 24-Sep-2017].
[205] A. Ng, “Supervised Learning - CS229 Lecture notes.” Stanford University, 2016.
[206] A. Ng, “Features and Polynomial Regression - Linear Regression with Multiple Variables,” *Coursera*. [Online]. Available: https://www.coursera.org/learn/machine-learning/lecture/Rqgfz/features-and-polynomial-regression. [Accessed: 21-Apr-2018].
[207] Scikit-learn Developers, “Nearest Neighbors,” 2016. [Online]. Available: http://scikit-learn.org/stable/modules/neighbors.html. [Accessed: 15-Jun-2017].
[208] Analytics Vidhya Content Team, “A Complete Tutorial on Tree Based Modeling from Scratch,” *Analytics Vidhya*, 12-Apr-2016. [Online]. Available: https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/#one. [Accessed: 14-Jun-2017].
[209] A. Ng, “Support Vector Machines - CS229 Lecture notes.” Stanford University, 2014.

[210] E. Frank, L. Trigg, G. Holmes, and I. H. Witten, “Technical note: Naive Bayes for regression,” *Machine Learning*, vol. 41, no. 1, pp. 5–25, 2000.
[211] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. Cambridge, Mass: MIT Press, 2006.
[212] M. A. Nielsen, “Using neural nets to recognize handwritten digits,” in *Neural Networks and Deep Learning*, 2015.
[213] N. K. Ahmed, A. F. Atiya, N. E. Gayar, and H. El-shishiny, “An Empirical Comparison of Machine Learning Models for Time Series Forecasting.”
[214] Mathworks UK, “Levenberg-Marquardt backpropagation - MATLAB trainlm,” 2016. [Online]. Available: http://uk.mathworks.com/help/nnet/ref/trainlm.html. [Accessed: 20-Mar-2017].
[215] Artificial Intelligence Courses, *1. Why it Helps to Combine Models*. 2013.
[216] A. Ng, “Principal Component Analysis - CS229 Lecture notes.” Stanford University, 2016.
[217] Mathworks UK, “Principal component analysis of raw data - MATLAB pca,” 2016. [Online]. Available: https://uk.mathworks.com/help/stats/pca.html. [Accessed: 21-Mar-2017].
[218] A. Courville, “Bayesian Methods for Neural Networks.” [Online]. Available: http://www.cs.cmu.edu/afs/cs/academic/class/15782-f06/slides/bayesian.pdf. [Accessed: 09-Sep-2015].
[219] R. M. Neal, “Bayesian learning for neural networks,” University of Toronto, 1995.
[220] M. Słoński, “Bayesian neural networks and gaussian processes in identification of concrete properties,” *Computer Assisted Mechanics and Engineering Sciences*, vol. 18, no. 4, pp. 291–302, 2011.
[221] D. J. MacKay, “Bayesian Interpolation,” *Neural Computation*, vol. 4, pp. 415–447, 1992.
[222] F. D. Foresee and M. T. Hagan, “Gauss-Newton approximation to Bayesian learning,” in *Neural Networks, 1997., International Conference on*, 1997, vol. 3, pp. 1930–1935.
[223] C. M. Bishop, “Bayesian Neural Networks,” in *Pattern Recognition and Machine Learning*, New York, NY, USA: Springer, 2006, pp. 277–281.
[224] D. Toulson, “The Use of Bayesian Inference in the Training and Interpretation of Neural Network Predictors,” Intelligent Financial Systems Limited, IFS-TR-02, 1997.
[225] Mathworks UK, “Adapt neural network to data as it is simulated - MATLAB adapt.” [Online]. Available: https://uk.mathworks.com/help/nnet/ref/adapt.html. [Accessed: 20-Mar-2017].
[226] MathWorks United Kingdom, “Gradient Descent with Momentum,” 2016. [Online]. Available: https://uk.mathworks.com/help/nnet/ref/traingdm.html. [Accessed: 21-Mar-2017].
[227] S. Naylor, M. Gillott, and E. Cooper, “A Post-Occupancy Case Study on the Relationship between Domestic Energy Use and Occupancy Profiles,” in *Sustainable energy for a resilient future*, 2016, vol. III, pp. 102–112.
[228] T. R. Whiffen *et al.*, “A concept review of power line communication in building energy management systems for the small to medium sized non-domestic built environment,” *Renewable and Sustainable Energy Reviews*, 2016.
[229] S. Naylor and M. Gillott, “Bluetooth-based Mobile Device Detection for Improved Energy-Saving Controls Systems,” presented at the 14th International Conference on Sustainable Energy Technologies, 2017.
[230] S. Naylor, M. Gillott, and G. Herries, “The Development of Occupancy Monitoring for Removing Uncertainty within Building Energy Management Systems,” presented at the 7th International Conference on Localisation and GNSS, 2017.

10 APPENDICES

10.1 Appendix – Green Street House Diagrams

a) Second Floor

b) First Floor

c) Ground Floor

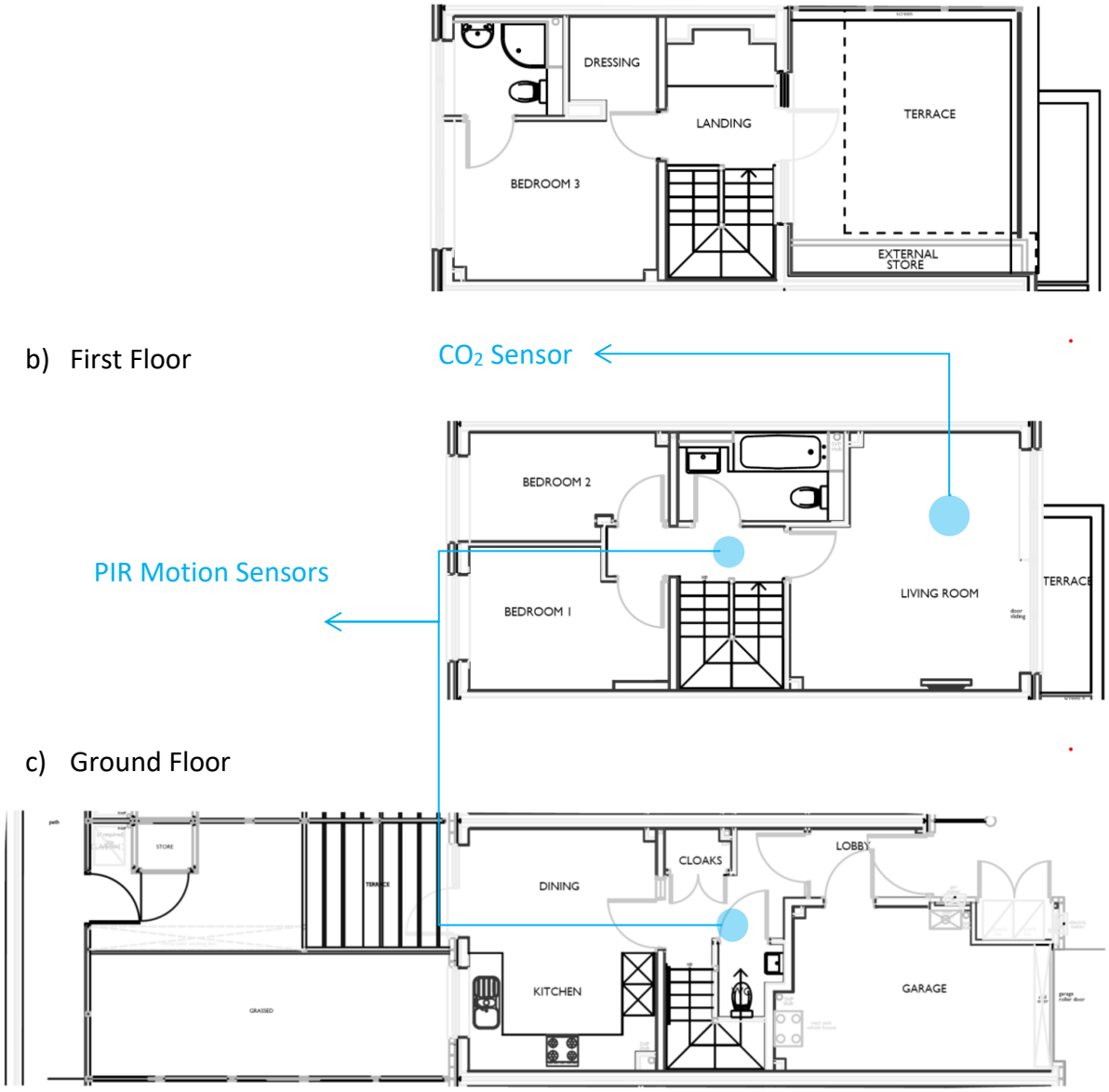


Figure 10-1 - Plans of house C showing the location of PIR and CO_2 sensors on a) Second Floor b) First Floor c) Ground Floor

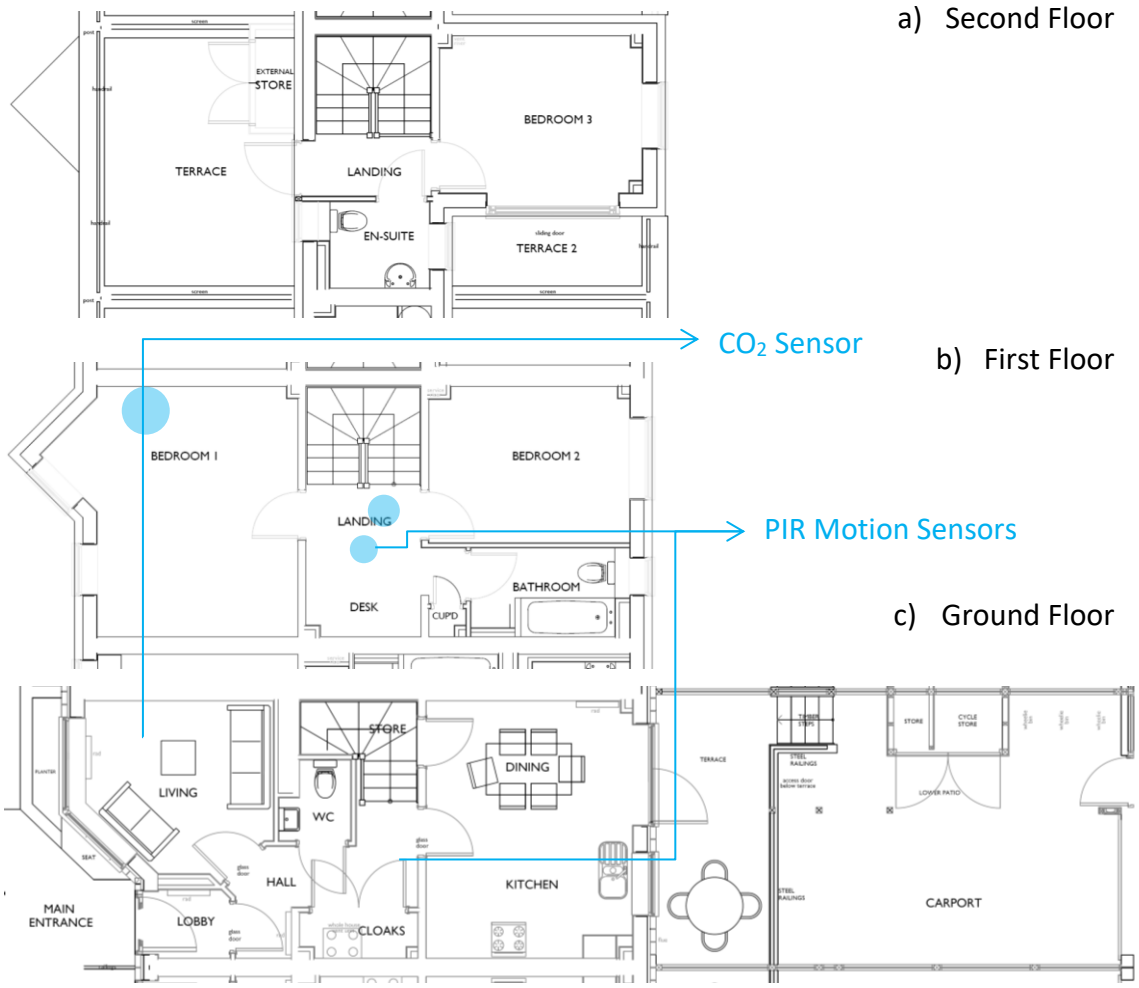


Figure 10-2 - Plans of house G showing the location of PIR and CO_2 sensors on a) Second Floor b) First Floor c) Ground Floor

10.2 Appendix – Explore Innovation Park Diagrams

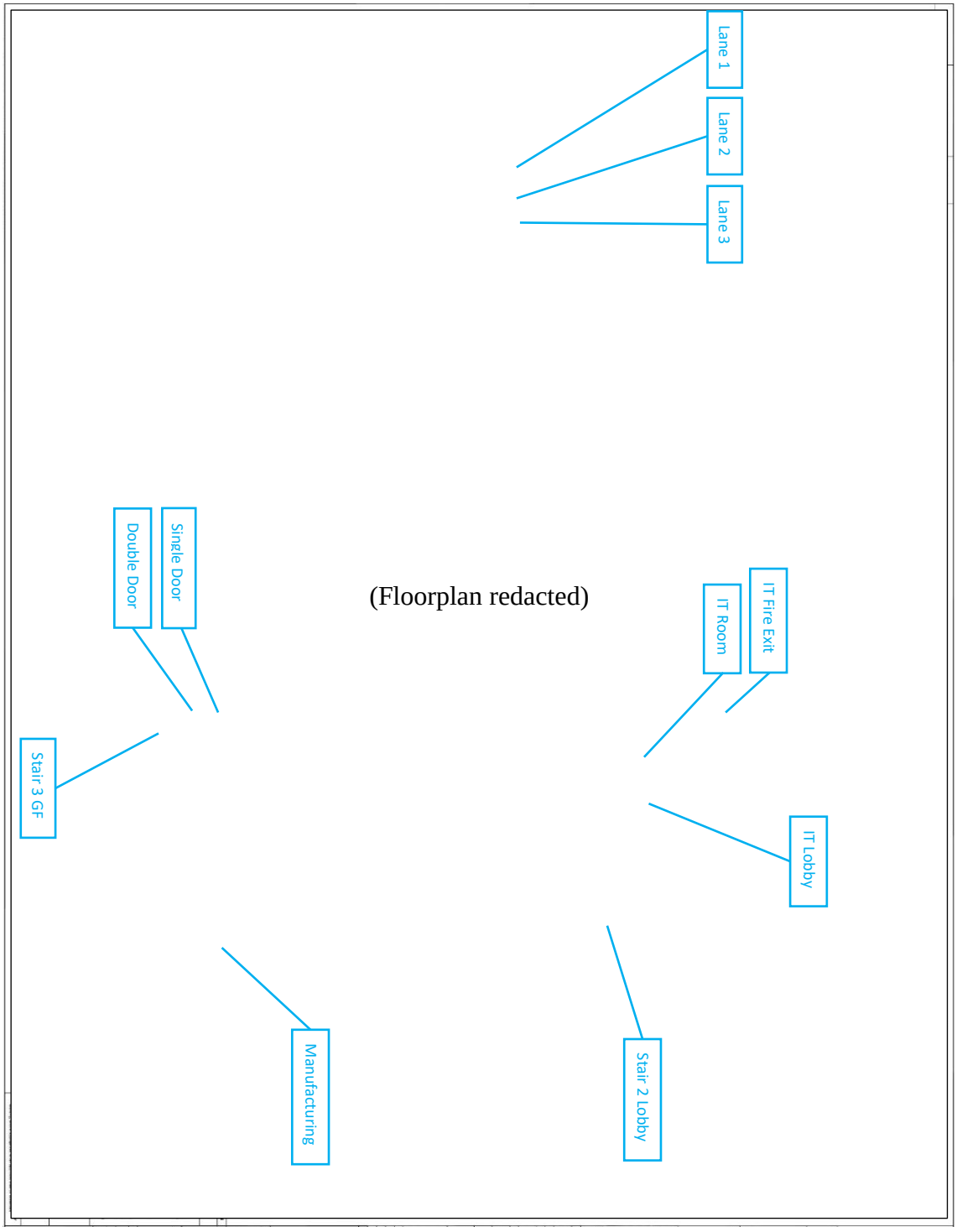


Figure 10-3 – Large office Case Study Layout: Ground Floor

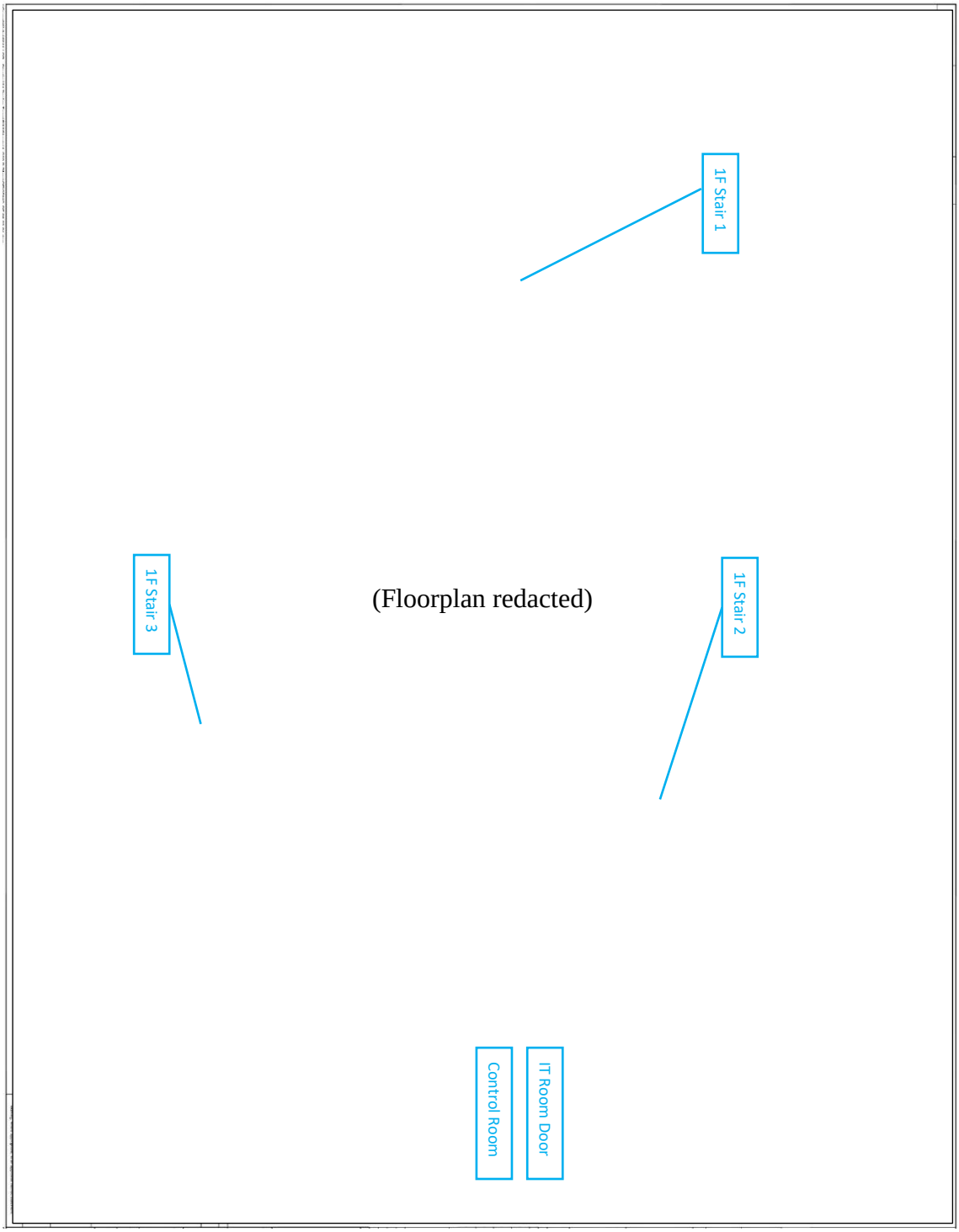


Figure 10-4 – Large office Case Study Layout: First Floor

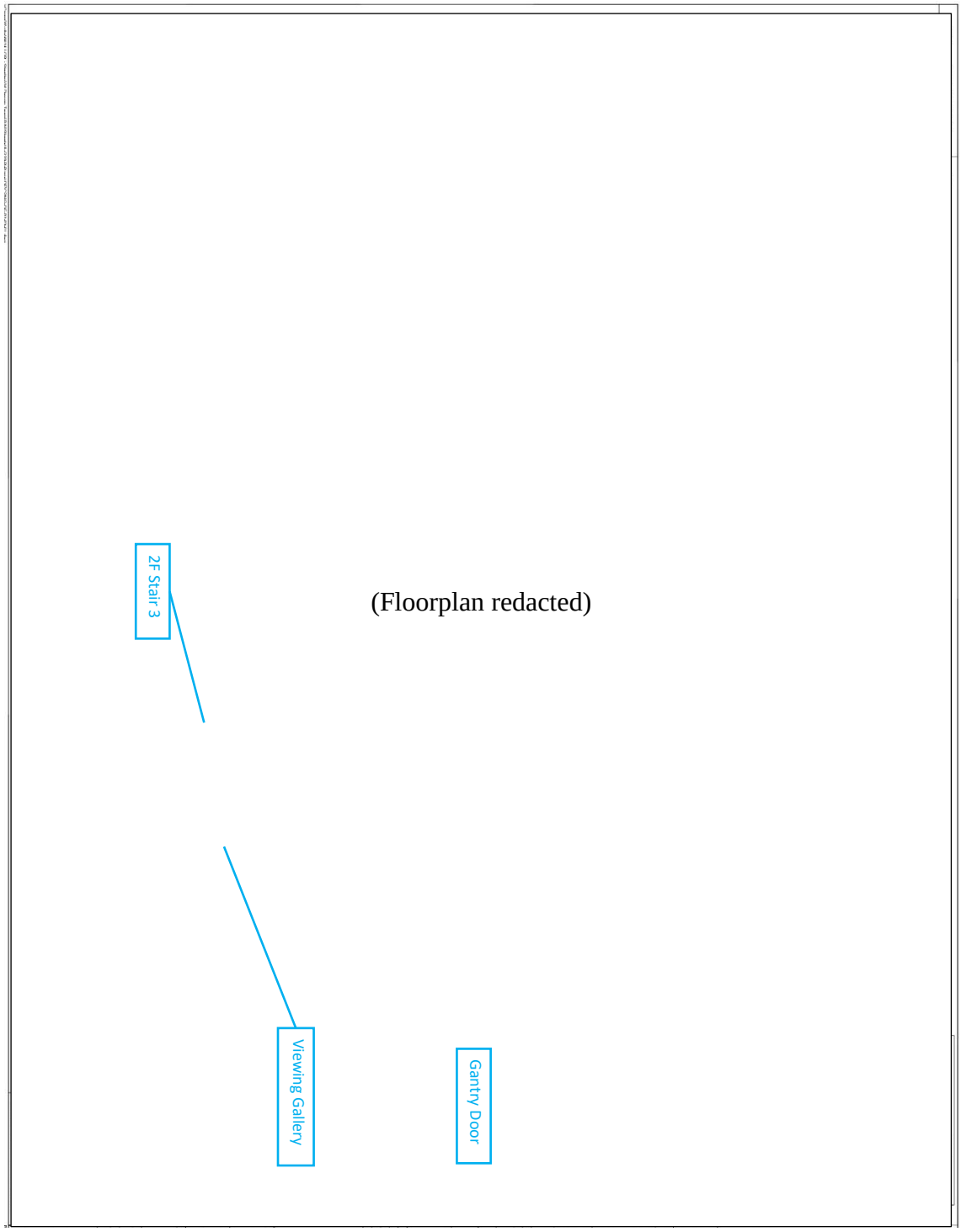


Figure 10-5 – Large office Case Study Layout: Second Floor

10.3 Appendix – Example Named Participant Consent Form

EXAMPLE PARTICIPANT INFORMATION SHEET

My project will be looking at the ways occupant data can be used to inform the controls in a building. I am using the existing sensing platform at the Mark Group House to collect environmental data on the temperature, CO2 level and motion in the office spaces. I intend to combine this non-identifying data with more explicit data on when the regular members of the office are present. This data will involve either or both of the following:

- Detection of a Wi-fi enabled device, selected by you for use in this study, while in the office
- Detection to the level of which room you are in while in the office through Bluetooth sensors placed around the house. This option is only possible for office members with handsets enabled with Bluetooth 4.0.

In order to test the accuracy of occupancy detection through these methods, I will also be running a test period in which I will ask office members to manually sign in/out of the building on paper sheets placed near the doors.

The above should involve a few minutes of your time for setup, and will otherwise be automatically collected upon your entry to the building. Those participating in the Bluetooth location sensing will be required to send their collected log file to the researcher periodically.

All data will be anonymised as far as possible before publication: no names or personal device identities will be associated with the data at the time of public presentation. Raw data will be stored on the data collection hardware and researcher's work computer(s), and will only be accessible by password entry. Note that the raw data alone will not identify the participants by name: data linking device IDs to real people will be stored in a separate file, also password protected. The raw data will only be accessed by the researcher and project supervisors.

Please note that participation in this research is completely voluntary. Participants are at liberty to withdraw at any time without any consequences. In accordance with university ethics policy, participants have a right to access data stored on them at any time.

Contact Details:

Sophie Naylor - _____

Supervisor Prof. Mark Gillott - _____

PARTICIPANT CONSENT FORM

Project title: Managing the uncertainty of occupant behaviour through real-time building energy evaluation and management

Researcher's name: Sophie Naylor

Supervisors' names: Mark Gillott, Ed Cooper

(Please tick all that apply)

- ☐ I have read and understood the statement of intent to use occupancy data from the Mark Group House offices.
- ☐ I have been given an opportunity to ask questions.
- ☐ I understand I can withdraw from the research project at any time without any consequences.
- ☐ I understand that while information gained during the study may be published, information which might potentially identify me will not be made public.
- ☐ I understand that I may contact the researcher or supervisor if I require further information about the research.

I consent to the use of my presence data in the Mark Group House via:

- ☐ Wi-Fi – nominated MAC address is ……
- ☐ Bluetooth
- ☐ Manual sign-in test period

Signed ……

Print name …… **Date** ……

Contact Details:

Sophie Naylor - _____

Supervisor Prof. Mark Gillott - _____

10.4 Appendix – Example Manual Location Data Collection Sheet

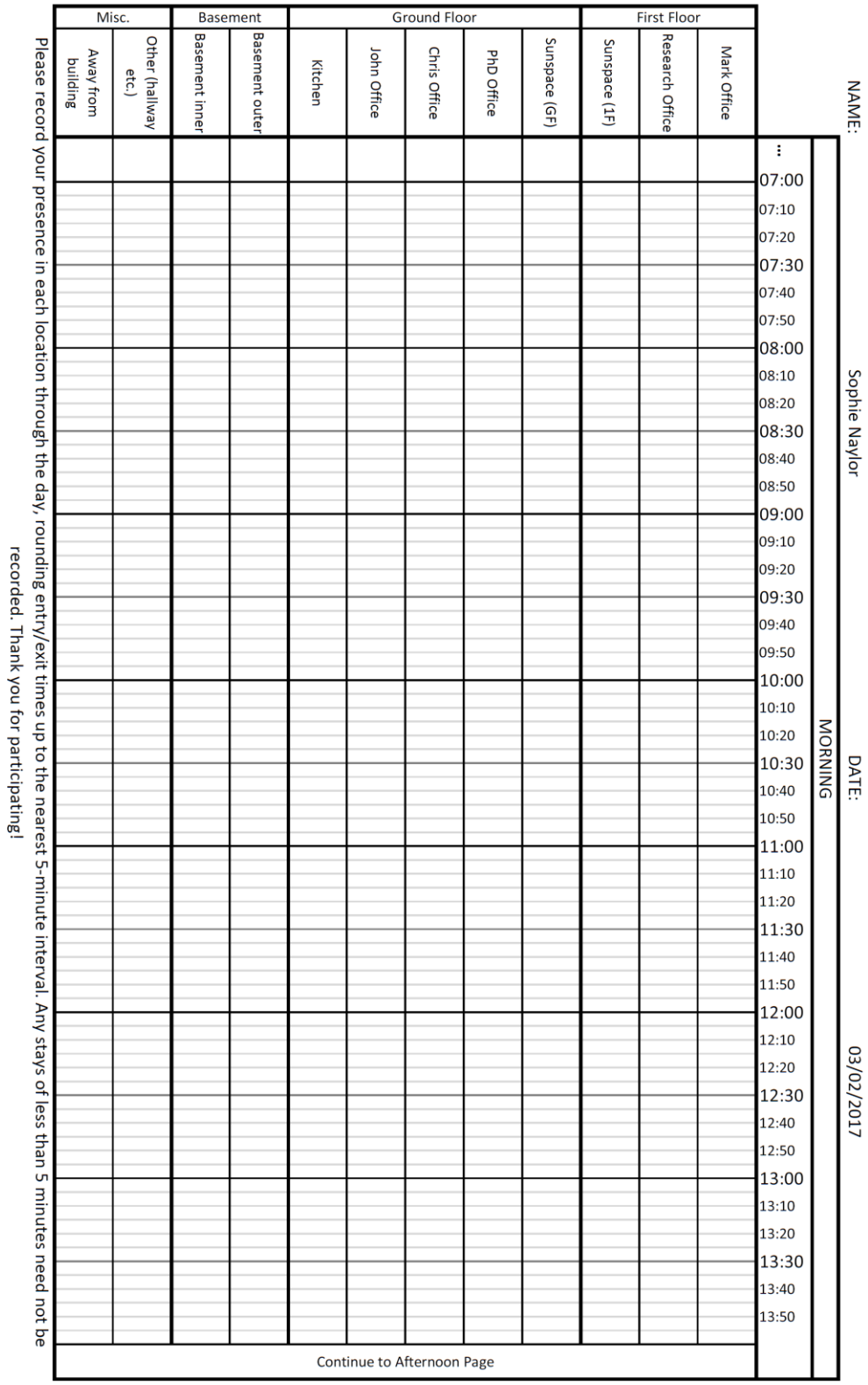


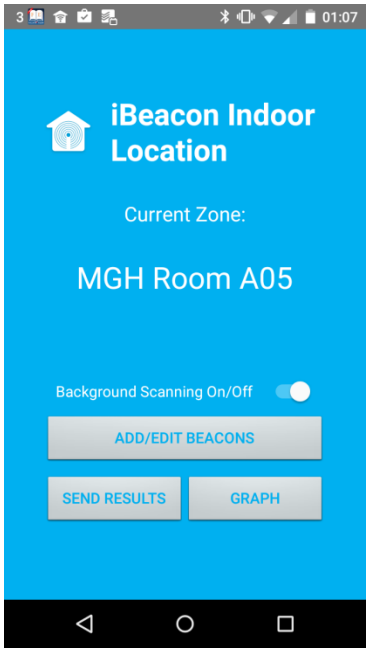
Figure 10-6 – Sample Manual Location Data Collection Sheet

10.5 Appendix – iBeacon Software Plain-Language Algorithms

10.5.1 Main Screen

When the screen is first opened:
Initiate receiver for changes in occupied status

Check whether scanning is enabled/disabled
Set scanning slider to enabled/disabled



Initiate listener for slider. When slider is changed:
Start/End background service
Save new scanning status to internal storage
Set display text to indicate status

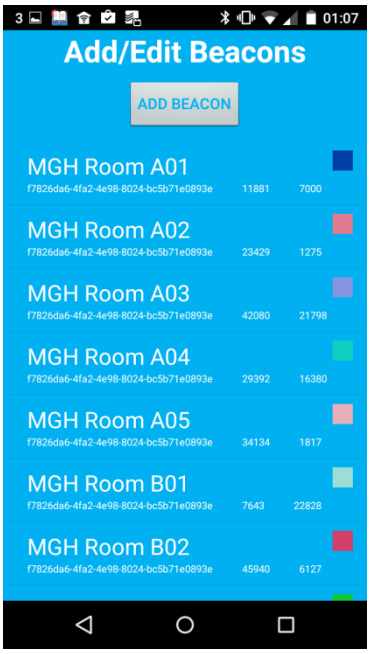
Initiate listener for 'Send Results' Button. When clicked:
Build email intent with log file attached
Initiate listener for 'Edit Beacons' Button. When clicked:
Start 'Edit Beacons' activity
Initiate listener for 'Graph' Button. When clicked:
Start 'Graph' activity

When the screen is started/restarted:
Load the list of known beacons from internal storage

Load the status of each known beacon from shared preferences
Update display to current beacon if occupied

When the screen is closed:
Stop receiver for changes in occupied status

10.5.2 Adding/Editing Beacon List



When the screen is first opened:
Initiate receiver for changes in occupied status

Load the list of known beacons from internal storage
Display the known beacons list

Initiate listener for 'Add Beacon Button. When clicked:
Build alert dialog for input of Name, UUID, Major, Minor
When dialog is entered:
Extract text from fields
If entered text is an appropriate format:
Add new beacon to the known beacon list
Save known beacon list to internal storage
Make a new beacon status in shared preferences
Refresh display of beacon list

```
When a beacon on the list is clicked:
        Build alert dialog for editing of Name, UUID, Major, Minor
        When edited text is entered:
                If entered text is an appropriate format:
                        Remove old beacon from beacon list, beacon status
                        Refresh display of beacon list
        When delete button is pressed:
                Build confirmation dialog
                If confirmed:
                        Remove beacon from beacon list, beacon status
                        Refresh display of beacon list
```

10.5.3 Background Service

```
Import altbeacon library classes
Import various android library classes

Initialise local variables
Set hardcoded scan duration and timeout duration

When service is first started:
        Initialise broadcaster for status change receiveron main screen

        Initialise manager for scanning service from altbeacon library
                Set scan duration and time between scans
                Set format for parsing beacon information to ibeacon format

        Build android system notification to denote background service is
working

        Load time of last activity from internal storage
                If last activity was more than 5 min ago
                        Log service outage
                        Exit any zones that are logged as occupied

        Update activity log

When service is closed:
        Update status of all beacons to zero

When altbeacon manager is started:
        Load list of known beacons from internal storage

        Loop each of the known beacons, identify any IDs common to them all

        Initiate altbeacon Range Notifier (listener for nearby beacons and
their distance). When a beacon is ranged:
                Update activity log time
                Check against last activity, log any service outage

                Start ScanTimer: a timer for the duration of the scan
period
                Load the list of known beacons from internal storage

                Put all beacons ranged into a set

                Loop the list of known beacons:
                        If the known beacon has been ranged:
                                Keep known beacon in the ranged list
                                Update distance if closer than a previous
entry
                Discard any beacons not in the known beacon list
```

```
                When ScanTimer finishes:
                        Get the list of beacons seen during the ScanTimer
                        Start or refresh TimeoutTimer: it resets beacons to
unoccupied if none are seen for the allotted time
                        Load the list of known beacons from internal
storage
                        Load the beacon status from shared preferences
                        Loop through the list of beacons seen during the
scan
                                Get the calculated distance, update
closest beacon
                        If no beacons are currently occupied:
                                Write entry of closest beacon to log file
on external storage
                                Send broadcast to update the display on
main screen
                                Update the beacon status in shared
preferences
                        Else:
                                If the closest beacon was not already
occupied:

                                        Write exit of last beacon to log
file on external storage
                                        Write entry of closest beacon to
log file on external storage
                                        Send broadcast to update the
display on main screen
                                        Update the beacon status in shared
preferences



                When TimeoutTimer finishes:
                        Update beacon status in shared preferences to sign
out of all zones
```

10.6 Appendix – Long-term Occupancy Dataset Generation Algorithm

Each occupant is assigned a main zone and lunchbreak zone.

One zone is designated as a meeting/lunch space. One zone is assigned as the visitor space. Teaching staff are assigned a single-occupancy zone from the remaining zones. Other office members are assigned a zone from the final remaining shared space zones. Each occupant is assigned either lunch zone or absence for lunch break period.

Each occupant is assigned a personal start/end time.

Start/end times are drawn from a Gaussian distribution, with mean of the base office worker profile as specified above, standard deviation determined by profile type.

Each occupant is assigned a personal variability level for start/end times.

Variabilities are drawn from a Gaussian distribution, with mean of 0, standard deviation determined by profile type.

Each occupant is assigned a likelihood of occupancy at weekends.

Drawn from random number generator determined by profile type.

Each occupant is assigned a profile of regular daily/weekly/monthly events.

Drawn from random number generator determined by profile type. Zone assigned per event, biased towards leaving the building.

Each occupant is assigned a likelihood of random day-long and sub-day absences.

Drawn from random number generator determined by profile type.

Generate Data:

Assign a start date based on the day of data generation

If longterm changes, generate days on which schedule and office changes will happen

Loop number of days of data required

Increment date

If longterm changes, check if current day on list. Generate new attributes for the relevant occupant agents and replace old attributes.

If new year, or profiles do not exist, Generate long-term absence profiles

Draw occupant start/end times from Gaussian, mean and standard deviation from personal profiles

Generate random sub-day events per occupant using personal absence likelihood and random event flag

If day is not in assigned long-term absence periods, Assign each person to their main zone according to their generated profile. Assign to lunch/absence zone for lunch break according to personal profile.

Apply regular daily/weekly/monthly patterns per occupant, according to profiles

Apply random events as generated above

If day is not in assigned long-term absence periods, Generate number of unknown visitor events that day from Gaussian, mean 2, std dev 1.2

Draw the number of people per visit from exponential distribution, lambda 1.5. Skew towards lower numbers by dividing any <4 by 1.5. Round up to whole numbers.

Generate start times with a random number between 10:00 and 16:00

Draw durations from exponential distribution, lam 0.5 h

Loop each event

Assign groups over 7 to the visitor space. Assign groups over 2 to the meeting space. Assign groups <=2 to a random zone, if that zone is occupied by a regular occupant

Next event.

Next day

10.7 Appendix – Attached CD containing Code Used in Thesis