Imputing Missing Occupancy Data Based on the Correlation of Seat Occupancy and Sound Level:

Taking Bartlett Library as An Example

Yuanru Gao

A Dissertation submitted in part fulfilment of the Degree of Master of Science Urban Spatial Science

at Centre for Advanced Spatial Analysis,

Bartlett Faculty of the Built Environment,

University College London

Submitted: 25 August 2023

Supervisor: Dr Valerio Signorelli

Module ID: CASA0010

Student ID: 22227790

Word Count: 11862 words

Abstract

In the evolving digital environment, libraries are undergoing a transformative journey fuelled by rapid technological advances. This transformation has given rise to the concept of smart libraries, where the fusion of cutting-edge technologies such as Artificial Intelligence (AI), data mining and the Internet of Things (IoT) have redefined library functions and services. By focusing on the integration of IoT technologies into libraries (using the Bartlett Library as an example), this paper explores the trends in occupancy and sound level over time within the Bartlett Library and the relationship between them, on the basis of which missing occupancy data can be predicted.

Heatmap, time series analysis and correlation analysis are used to depict seat occupancy patterns and noise level in the library, including their correlation and trends. Grasping these trends and correlations is pivotal for addressing missing occupancy data. For predicting missing occupancy, deep learning models LSTM and GRU are employed. The GRU model outperforms LSTM in accuracy, though there is still potential for improvement.

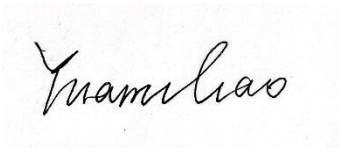
The broader significance of this research lies in the application of IoT technologies to enhance library management and user experience. By revealing the intricate interactions between occupancy and sound level, the study informs decision-making and paves the way for future optimization. In addition, the methodology employed provides ideas and validation applicable to similar building spaces, while suggestions for improving the measurement and optimisation models suggest future research

directions. In conclusion, this study exemplifies the potential of IoT being used in smart libraries and provides a blueprint for future research in similar contexts.

Contents

Declaration of Authorship

I, Yuanru Gao, hereby declare that this dissertation is all my own original work and that all sources have been acknowledged. The dissertation is 11862 words in length from the introduction to conclusion inclusive, excluding footnotes, tables and figures. Word count by Microsoft Word.



Signed: _____

Date: 24 August 2023

Tables

Figures

Abbreviations

ADF: Augmented Dickey-Fuller

AI: Artificial Intelligence

ARIMA: AutoRegressive Integrated Moving Average

ARMA: Autoregressive Moving Average

BAS: Building Automation Systems

BMS: Building Management Systems

dBA/Db(A): A-weighted decibel

EDA: Exploratory Data Analysis

GRU: Gated Recurrent Unit

KNN: K-Nearest Neighbors

LSTM: Long short-term memory

MA: Moving Average

MAE: Mean Absolute Error

MSE: Mean Squared Error

PIR: Passive Infrared

RF: Random Forest

R2: R-squared

RNN: Recurrent Neural Networks

RMSE: Root Mean Squared Error

S.D.: Standard Deviation

Sound Pressure Level: SPL

SVM: Support Vector Machines

Acknowledgement

First, I wish to express my sincere gratitude to my supervisor Dr Valerio Signorelli. During the period of completing my dissertation, Valerio has been very patient in answering my questions, and his positive responses have been a great motivation for me to be able to complete my dissertation. I also gained a lot of great ideas and perspectives by talking with Valerio. It is my luckiest and happiest thing to be supervised by him and to discuss with him. I sincerely wish Valerio a wonderful and lovely life and career ahead.

I would also like to thank Sarah Turk, the manager of the Bartlett Library, for providing me with so much information about the Bartlett Library.

I appreciate that I was able to study at UCL CASA this year and all the teachers here are extremely nice. I have broadened my research horizons, learnt what I wanted to learn and met many amazing students here. I am grateful to all of you for taking care of me this year, and I would like to thank every staff and student at CASA.

Last but not least, I would like to thank my family. A lot of things happened at home last year and it was a very difficult stage, but my family supported my decision without any hesitation, especially my parents, who are my strongest support no matter what. I thank them for their understanding and support.

This paper is dedicated to all the people I mentioned above as well as to myself. Thank you all for being my North Star in my life journey.

May all those whom I cherish, care for, and respect could have a life as beautiful as poetry.

1 Introduction

1.1 Context and Motivation

In the digital era, libraries are undergoing a profound transformation in the face of rapid technological advances(Jadhav and Shenoy, 2020). In particular, when cutting-edge technologies such as Artificial Intelligence (AI), data mining and the Internet of Things (IoT) are combined with library services, they give libraries new definitions and functions. The concept of the smart library is gradually emerging. It not only provides users with richer and more convenient resources but also greatly improves the operational efficiency of libraries. IoT technology plays a key role in it. Its core value lies not only in the sensors on the objects but more importantly in its ability to automatically track and share information, transforming physical objects into intelligent virtual objects that can be interactive (Madakam, Ramaswamy and Tripathi, 2015; Liang, 2018).

Acoustic studies play an important role in the library environment. Firstly, libraries and study spaces, due to their specific functionality, usually set clear noise level requirements with the aim of creating a peaceful study and reading environment for the users, and acoustic studies in such scenarios are concerned with the comfort of the users. Secondly, sound level data can also be seen as a powerful indicator for assessing library usage and estimating the flow of people or occupancy information. Noise information and occupancy information are particularly important for library users. Occupancy information can help them to know whether there are vacant or private locations; noise information provides an indicator of the study environment that

users can refer to, so that they can decide whether to study in a particular area or not, based on their own sensitivity to and preference for noise. However, the large amount of data collection also brings challenges, and how to deal with the missing data problem that often occurs on the Internet of Things is an urgent problem that needs to be solved nowadays.

Looking across the smart city landscape, the combined application of IoT and other innovative technologies is transforming data into valuable information and knowledge. This not only provides cities and public facilities with the opportunity to reframe and optimise their services but also brings about a more efficient and intelligent service experience for the community. Libraries, as an important public facility, with the help of IoT technologies, can not only better connect their resources and services, but also establish closer ties with their users, and moreover, intersect with other smart building applications to generate new opportunities.

1.2 Research Questions

Based on the above background, this study intends to provide a deeper understanding of the spatial distribution and temporal trends of seating occupancy and sound level at Bartlett Library, as well as the correlation between them. Considering the continuity and large-scale missing data, this study will attempt to use deep learning models to predict the missing data. Specifically, this study aims to answer the following questions:

1. How is the spatial distribution of seat occupancy and how is noise condition in the Bartlett Library during the research period?
2. What are the temporal trends of sound level and seat occupancy in the Bartlett Library during the study period?
3. Is there any correlation between seat occupancy and sound level at the Bartlett Library?
4. How to impute the missing occupancy data for the Bartlett Library with the assistance of the relationship between occupancy and sound level?

1.3 Study Structure

The rest of the document is structured as follows:

- Chapter 2: Literature Review

A comprehensive review of the relevant literature related to the research topic and associated methodology is carried out.

- Chapter 3: Methodology

It primarily describes a research process based on time series analysis, including data sources, collection, descriptive statistics, sensor installation, noise level and occupancy analyses. In addition, the methodology addresses the principles of

exploratory data analysis, correlation analysis, and two predictive models (LSTM and GRU). Finally, there is an ethical consideration.

- Chapter 4: Result and Analysis

Spatial patterns of library use, temporal trends, correlation and prediction value are the results derived using the methods in the methodology to address the four research questions.

- Chapter 5: Discussion

Interpret the above results and list the limitations and prospects of this study.

- Chapter 6: Conclusion

Outline the purpose and methodology of the study, research findings, and the importance of the study.

- Appendix

2 Literature Review

There have been a number of longitudinal and cross-sectional studies that have explored the effect and connection of seat occupancy and sound level data acquired by IoT technology in buildings, and these studies help to address the core questions of my current research: What are the trends in seat occupancy and sound level in Bartlett Library, what is the connection between them, and how to use this connection to help estimate missing seat occupancy data in the library. The next part reviews these relevant studies to reveal the key findings and remaining gaps in the existing research.

2.1 Techniques and Tools for Research within Buildings

Research within buildings focuses on indoor environments, comfort and health, interior design, human behaviour and activities. Researchers often advance these studies with the help of Building Management Systems (BMS), sometimes called Building Automation Systems (BAS) or Energy Management, a specific application of the Internet of Things (IoT). At its core, BMS is about integrating and controlling multiple systems within a building to meet the building's needs for comfort, availability, security, and energy management. In order to achieve these goals, BMS needs to continuously collect and analyse data from sensors, which helps to optimise energy use, monitor environmental parameters in real-time and manage utilities efficiently (Minoli, Sohraby and Occhiogrosso, 2017) Regarding sensors in BMS, Dong's team suggested that

they can be broadly classified into three main categories (Dong et al., 2019), as shown in the table below.

Table 1 Comparison of the Use and Performance of Three Types of Indoor Sensors

Name	*Sensor category*	*Application*	*Cost*	*Accuracy*	*Advantage*	*Disadvantage*
Image based sensors	Detecting Occupancy in The building environment	Image processing and recording: surveillance and security (Pedersen et al., 2016).	Medium-High	High	Clear visual verification; Wide range of usage scenarios	Privacy issues; Illumination conditions; Storage needs
Passive infrared (PIR) sensors		Indoor occupancy detection (Pedersen et al., 2017)	Low-Medium	Medium	Low power consumption; Easy installation; Reliable motion detection	Limited detection range; Unable to tell if it's human or not
Radio sensors		Security Systems, indoor occupancy detection (Liu *et al.*, 2020)	Medium	Medium	Can detect through walls; Have a long detection range.	Easily interferes with other equipment
Carbon dioxide sensor	Monitoring indoor building environment parameters	Carbon dioxide intensity for estimating indoor occupancy (Meyn *et al.*, 2009; Wei *et al.*, 2022)	Medium	Medium	Real-time monitoring; Full understanding of the indoor environment	Methods for estimating indoor population based on carbon dioxide often suffer from delays and heterogeneity (Wang, Chen and Hong, 2018); Complexity of setup and calibration
Sound level, illumination, humidity and temperature sensors		Multivariable combination to predict indoor occupancy (Ekwevugbe *et al.*, 2013; Ang, Dilys Salim and Hamilton, 2016)	Medium	Medium-High		
Wearable sensors	Other types of sensors	Understand the behaviour of people indoors (Dong *et al.*, 2019)	Medium-High	High	Real-time tracking; Detailed behavioural analysis	User privacy issues; Reliance on user wear
Smart Phones			High	High	Multi-functional; High user familiarity.	User privacy issues; Battery consumption issues

As can be learnt from the table above (Table 1), the use of sensors in research conducted indoors has many advantages. These advantages include high efficiency, continuous remote observation, and easy data transmission and storage. In addition, a significant advantage is the reduction of time-consuming activities. For example, Xia (Xia, 2005) used personal observation in data collection on library seat occupancy for a sustained period of three months, where observations and data collection were carried out for several hours each day except at night. However, It is worth noting that, while the use of sensors for remote observation reduces repetitive work, it does not allow for the simultaneous advantages of in situ observation and surveys, such as the fact that in situ observation and surveys allows for a more intuitive acquisition of data information with more detail, which is more flexible than sensors that can only detect a specified amount of data, as well as the fact that the in situ observation and surveys are more stable and are not as vulnerable as the sensors to technical malfunctions. Similarly, it is important to note that high-precision and informative sensors are often accompanied by issues of personal privacy risk. For more informative and consistent information, observations and surveys can be combined with sensor detection technique (Gilani and O'Brien, 2017).

From Table 1, we can also learn that when selecting sensors, it is important to consider their functionality, cost, privacy and accuracy, but it is difficult to combine the pros and cons of one type of sensor, so sensor fusion and data fusion can be used to compensate for the shortcomings between the sensors, which means that combining the data information from multiple sensors, which will result in a more accurate and

comprehensive interpretation of the data than using a single sensor. In this study, PIR and Sound Level sensors are used in combination, these two sensors are accurate, inexpensive and easy to obtain and install in the study area, so these two sensors are chosen to conduct an in-depth study of seat occupancy and sound level in the library.

After the sensors detect the data, they need to be transmitted to the BMS through wired or wireless means, and the users can get the data from the user terminal of the BMS. The process can be referred to Figure 1 and the specific methodology is referred to in (Chapter 3 Methodology). The data collected by the PIR in this study is transferred to the UCL API database and the user then calls the API to get the data, while the sound level is transferred to a specific address via LoRaWAN, and the user can log in and download the data.

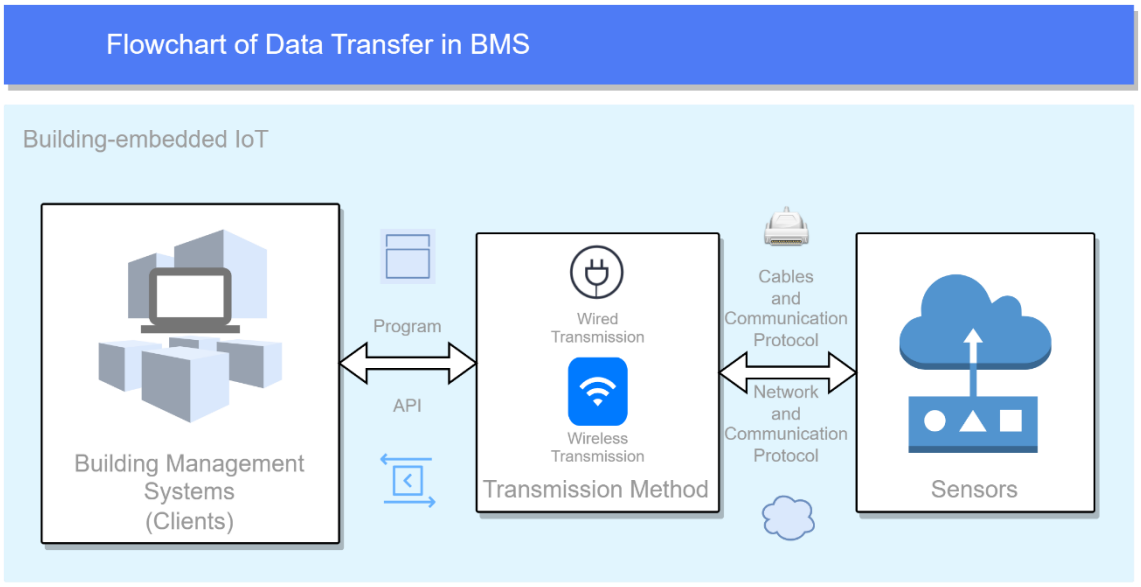


Figure 1 Flowchart of Data Transfer in Building Management Systems

In this subsection, I discussed techniques and tools for data acquisition in indoor research. In the next section, I am going to talk about the current state of research in indoor occupancy and sound level, respectively.

2.2 Library Use and Seat Occupancy Study

As an essential public place on campus or even in the urban area, the library provides students and residents with important academic resources and space for learning and communication, and its seat occupancy pattern and library management are directly related to the reasonable use of public resources. Especially during the academic peak period, how to effectively manage and optimise seat occupancy has become a prominent issue. Inappropriate occupancy management may lead to a waste of resources. Therefore, an in-depth study of occupancy patterns in libraries is of remarkable significance.

One of the focuses of this study is to longitudinally analyse the occupancy patterns of library seats using the occupancy data obtained from PIR sensors to gain insights into the actual use of the seats and the changes in occupancy patterns over time, which not only tests the performance of sensors as a tool for remotely monitoring, but also provides more scientific and reasonable suggestions for space planning and management in libraries to ensure efficient and balanced use of resources. Previous researchers have studied the occupancy patterns of libraries longitudinally: Wang, in his longitudinal study of the occupancy patterns of the University of Reading Library, used thermal imaging sensors to collect data, and concluded that the length of stay in

the library remained consistent throughout the academic year (including holidays), and that the peak occupancy period was basically stable between 3:30pm and 5:30pm (Wang, Patel and Shao, 2023). The thermal imaging sensor used in that study and the PIR sensor used in this paper have similar detection principles, both using infrared radiation and heat detection, but the thermal imaging sensor is good at large-range detection and can accurately display the temperature of each point in the detection range, which is conducive to obtaining information on the population density and movement in the entire space. PIR sensor, compared to the thermal imaging sensor, is mainly used to detect whether there is an object movement or not, which belongs to the small range observations, but with low cost and low power consumption.

In addition to this, through cross-sectional studies, researchers have also identified a variety of factors that influence seat occupancy patterns, including privacy (Cha and Kim, 2015), distance from other users (Tunahan and Altamirano, 2022), and connection to outdoor (Gou, Khoshbakht and Mahdoudi, 2018). Tunahan and Altamirano (2022) also chose Bartlett Library as their study area, and they used a non-invasive method, which was to use seat occupancy data stored in the OccupEye Cloud, to determine changes in the seating preference of Bartlett Library users after the COVID-19 Pandemic, and OccupEye is a sensor system for monitoring and analysing the usage of offices, libraries, meeting rooms and other places. They concluded that there was a decrease in seat usage after the COVID-19 Pandemic and a preference for individual desks or desks at a certain distance from other students at a distance from the desks. The methodology and conclusions of Tunahan and Altamiranos' study provided excellent ideas and prerequisites for this study's research into the occupancy

of the Bartlett Library. It is worth noting that other findings from the study on the correlation between library use and environmental factors revealed that a good environment is often preferred by students, greatly affecting their choice of seating (Keskin, Chen and Fotios, 2015). Therefore, the next subsection will discuss the indoor environmental factors study.

2.3 Study of Indoor Environmental Factors

2.3.1 Research on the Impact of Indoor Environmental Factors

Poor indoor environments affect negatively on people' health, comfort and productivity. A high-quality indoor environment not only enhances the user experience, but also helps to reduce building energy consumption (Fisk, 2000). In order to understand more specifically how the quality of indoor environments affects people's comfort, Frontczak and Wargocki divided it into four main areas: thermal comfort, visual comfort, acoustic comfort and indoor air quality (Frontczak and Wargocki, 2011).

Among these, thermal comfort is perceived most directly, so it is the first concern of building occupants. However, other factors also deserve deeper consideration by researchers because they have different aspects and levels of influence on building occupants' experience and building energy consumption. For example, visual comfort is closely related to interior light conditions and window layout. The team of Dong stated that sunlight and light brightness are key elements in determining visual comfort

(Dong et al., 2019). This was verified in an empirical study at Bartlett Library, where it was found that the main factor considered by students when choosing a seat was daylight, followed by privacy, outdoor views, and the level of quietness (Tunahan, 2021). In indoor air quality studies, monitoring of pollutant concentration and carbon dioxide concentration is usually used to reflect the environmental quality directly or indirectly (Sahu and Gurjar, 2021). Acoustic comfort studies in which respondents indicated that they preferred to study or work in an environment with low noise or comfortable sound (Zotoo et al.,2023).

2.3.2 Acoustic Factors in Indoor Space

Acoustic factors take an important place in the study of indoor environments. In order to delve deeper into the sub-field of acoustics research, we need to first identify two main acoustic expressions: Noise and Sound Pressure Level (SPL). Noise research is concerned with the impact of sound on the indoor environment and its occupants, while SPL focuses on detecting the effect of sound propagation in a room or as a reference variable to estimate the occupancy rate of the room.

Noise, in a broad sense, belongs to a branch of acoustic research, while in a narrower interpretation, it deals with the nature of noise and methods of control. Noise was defined by the Wilson Committee in 1963 as 'unwanted sound'(‘Wilson Committee Report on Noise’, 1963). However, the Noise Policy Statement for England takes a more positive view of noise as an unavoidable by-product of a vibrant society, at times exciting and at other times an unwelcome intrusion (Department for Environment,

2010). As mentioned in Chapter 2.3.1 Research on the Impact of Indoor Environmental Factors, researchers have been studying the impact of various environmental factors on office efficiency, comfort and energy consumption, including noise factors. One of the more representative studies examined the effects of seating and noise level on students' perceptions of the library environment, using the Jiangsu University Library as an example (Zotoo et al., 2023). By quantitatively analysing the results of the questionnaire, the study concluded that most respondents were satisfied with the noise level in the library. However, people's response to noise varies depending on a number of factors such as intensity, frequency, duration and personal sensitivity (Kim, 2015).

Sound Pressure Level is a physical quantity of sound that describes the change in pressure of a sound wave. If the general term refers to the intensity or loudness of sound, Sound Level can be used. As for sound level research, Hodgson and Nosal explored its relationship with classroom occupancy in 2002, focusing mainly on the effects of sound propagation in classrooms with audiences (Hodgson and Nosal, 2002). Furthermore, other researchers estimated indoor occupancy using data fusion of multiple sensors that included detection of sound level (Ekwevugbe, Brown and Fan, 2012; Ekwevugbe et al., 2013), but the current study found that using sound level as an indicator to estimate indoor occupancy was not as accurate as carbon dioxide content (Ang, Dilys Salim and Hamilton, 2016).

From the methods and results of the above respective studies, it can be understood that traditional environmental assessment methods, such as on-site observations or

questionnaires, are often strongly influenced by the subjective emotions of the respondents, which may lead to biased data. In addition, these methods are relatively demanding in terms of time and human resources. In particular, when using sound level as the only research indicator, its efficacy in estimating indoor occupancy is limited and often needs to be combined with other data to increase accuracy. Therefore, in order to bypass these problems, I decided to take an innovative approach: combining occupancy and sound level data to provide an in-depth exploration of occupancy patterns within libraries and the relationship between is approach not only reduces the difficulty of data collection and shortens the research period, but also improves the accuracy of the study and the applicability of the methodology. After establishing such a methodological system, it can be applied to other libraries or different regions, reducing case-specific limitations.

2.4 Time Series and Missing Data Estimation

Time series analysis is particularly appropriate for the long-term continuous noise and seat occupancy data collected in this study. In the following section, the current status of time series analysis and its use for estimating missing data is explored in detail.

Time series have a rich history and many models suitable for analysis, classification and forecasting have been developed. From a theoretical point of view, a time series is a collection of observations made in chronological order (Brockwell and Davis, 2009). Time series data are characterised by large volume, continuity and equidistance (Fu, 2011). Both occupancy and sound level data addressed in this paper meet these

criteria. The application of time series analysis focuses on two main aspects: identifying the intrinsic structure and patterns of the data; and forecasting the future data based on these patterns. Commonly used methods such as Moving Average (MA) and Autoregressive Moving Average (ARMA) are capable of handling non-stationary time series. As the application of IoT technology in indoor research is developing, time series analysis has also played a key role in building energy consumption prediction (Deb et al., 2017) and environmental data de-noising (Zhou et al., 2020).

However, this study encountered problems during data collection. This research uses the UCL API to access occupancy data. UCL API[1]is an API integration platform that can be used to access university data developed by a team of students at University College London. University College London students log in to their accounts to obtain a personal key and can use the platform's API commands to obtain all workspace sensor information with real-time and historical data on occupancy of seats detected by the sensors, as well as university information such as seat bookings, school departments, and timetables. UCL Assistant, as an official University College London application, utilises the UCL API to access the data.

Because of the update of the UCL API system, the occupancy data from 18th June to 9th July were missing, while the sound level data were collected normally. Faced with

[1] The address of the UCL API is: https://uclapi.com/

this missing data, I need to fine-tune my research strategy and find appropriate methods for data estimation.

In fact, missing data is a common problem in BMS. Past researchers have used a variety of methods to cope with it, such as:

1. Deleting/ Ignoring missing data: This is suitable for cases where the observation period is long, and the data are missing randomly. However, this method leads to a loss of information.

2. Imputation / Interpolation: Mean Imputation or K-Nearest Neighbors (KNN) Imputation can be used. It is suitable for time series data that are stable and are either randomly missing or transiently missing due to exogenous factors (Pourahmadi, 1989).

3. Model prediction methods: time series models, such as AutoRegressive Integrated Moving Average (ARIMA) and exponential smoothing, are suitable for cases where the time series pattern is simple; Machine Learning models, such as Random Forest (RF) and Support Vector Machines (SVM) (Camastra et al., 2022), are mostly used for classification prediction; Deep Learning models: Long short-term memory (LSTM) (Ma et al., 2020) and Gated Recurrent Unit (GRU) (Yamak, Yujian and Gadosey, 2019), for the case of complex time series patterns.

In particular, the method proposed by Chen et al (2021) for optimising LSTM models using transfer learning appears to be particularly effective for dealing with the missing problem of large-scale continuous data.

In summary, considering that the missing data in this study is large-scale and non-random, it is assumed that using deep learning models for estimation is a sensible choice.

2.5 Conclusion

This chapter explores in detail the techniques and tools used in research within buildings, in particular the use of sensor technology in building management systems, library seat occupancy studies, and indoor environmental factor studies, including acoustic factors. In the section on time series and missing data estimation, I mentioned the challenges in data collection and different approaches on how to address them. However, the obvious research gaps in this in-depth series are how to combine different types of sensor data, to address the problem of missing data, and how to optimise and estimate large-scale continuous data with deep learning models. Although many research methods and techniques have been mentioned and validated, how to ensure the feasibility and accuracy of these methods in practical applications is still a matter of investigation and discussion.

3 Methodology

3.1 Introduction

Based on an in-depth analysis of the existing literature (Chapter 2 Literature Review), the following method is designed to address the aforementioned problem in this study. This chapter will describe the implementation steps and principles of the method in detail.

3.1.1 Case study design and approach

Case study research is an in-depth exploration and investigation into the complexity and uniqueness of specific projects or systems within real-life settings. It should not be regarded as a singular research method but rather as a research design framework that can incorporate various methodologies (Denzin and Lincoln, 2005; Thomas, 2011).

For this study, I chose the Bartlett Library as the subject of the case study based on several key considerations. Firstly, as a student at University College London, I can conveniently communicate and coordinate with the university department and the library, facilitating the installation of sensors and data collection. Secondly, the Bartlett Library, although compact in its spatial dimensions, has dense human traffic and is managed efficiently by staff. This means that while its space is limited, the data derived is rich and stable. Finally, in the library, each study desk is equipped with a detector, simplifying the process of collecting seat occupancy data.

Employing a quantitative analytical approach, I delved into the relationship between seat occupancy data and sound level in the library. This specific case offers valuable insights into the interplay between occupancy and the environmental conditions of a library.

To ensure the reliability and validity of our study, I took several precautionary measures. This included employing standardized procedures for data processing and storage to ensure consistency and accuracy. Additionally, all research activities adhered to ethical guidelines and received appropriate ethical review. Moreover, the analytical code used in this research has been made publicly available on GitHub (Gao, 2023) for other researchers' verification and use.

3.1.2 Workflow

As shown in the Figure 2, the workflow of this study is divided into 4 parts which are Fundamental work, Data Preparation, Analysis and Prediction.

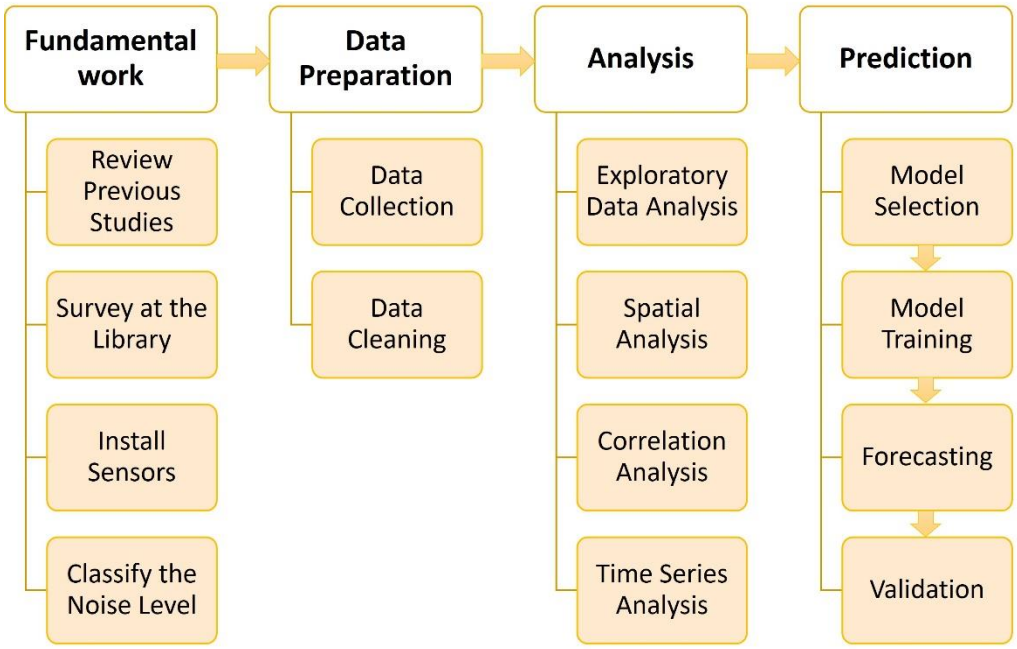


Figure 2 Workflow

3.2 Data Source

The main data information, format, source, size and time period of this study are listed in the table below (Table 2).

Table 2 Details of Data

Data	*Data Information*	*Size*	*Time Period*	*Source*	*Storage Format*
Occupancy and sound level data	Number, Room ID, Date, Occupancy counts (Column name: count_1s), Sound Level (Column name: Value)	1585 rows; 1154 rows; 1513 rows	2022/06/18-2022/07/09 (Only occupancy data) 2023/06/02-2023/06/17 AND 2023/07/10-2023/07/30 (Both occupancy and sound level)	Occupancy: UCL API: workspaces/historical/data; Sound level: Installed sound level sensor data-collection software	Occupancy: JSON-> CSV; Sound level: CSV
L Aeq data	Date, Room ID, L_Aeq Value	177 rows	2023/06/02 - 2023/07/30	Calculated sound level data	CSV
Base map of Bartlett Library	2D Map of Bartlett Library	3,649 KB		UCL API: workspaces/images/map	PNG-> TIFF
Bartlett Library Geographic data	Sensor ID, X coordinate, Y coordinate	11 KB		UCL API: workspaces/sensors	JSON-> GEOJSON

Note: In parentheses is descriptive information such as column names or time periods corresponding to the data

3.2.1 Study Area

This research was carried out in the Bartlett Library at University College London, situated on the ground floor of a six-story building named Central House, in the centre of London. The building is oriented east-west and faces a busy street, which is in close proximity to both a hospital and a university. Due to the regular flow of vehicles and frequent passing of ambulances, the surrounding area is bustling and noisy. As illustrated in Figure 3, the library consists of three main areas. Room 1 is a smaller group study area with 8 shared office desks and 4 partitioned study desks. Room 2 is a spacious open study space with a skylight, accommodating 32 shared study desks.

Room 3 serves as the library's open shelf area, offering 11 study desks and several side windows facing north and east. The library opens from 9 a.m. to 8 p.m. from Monday to Friday. On Saturdays, it's open from 11 a.m. to 6 p.m. and remains closed on Sundays.

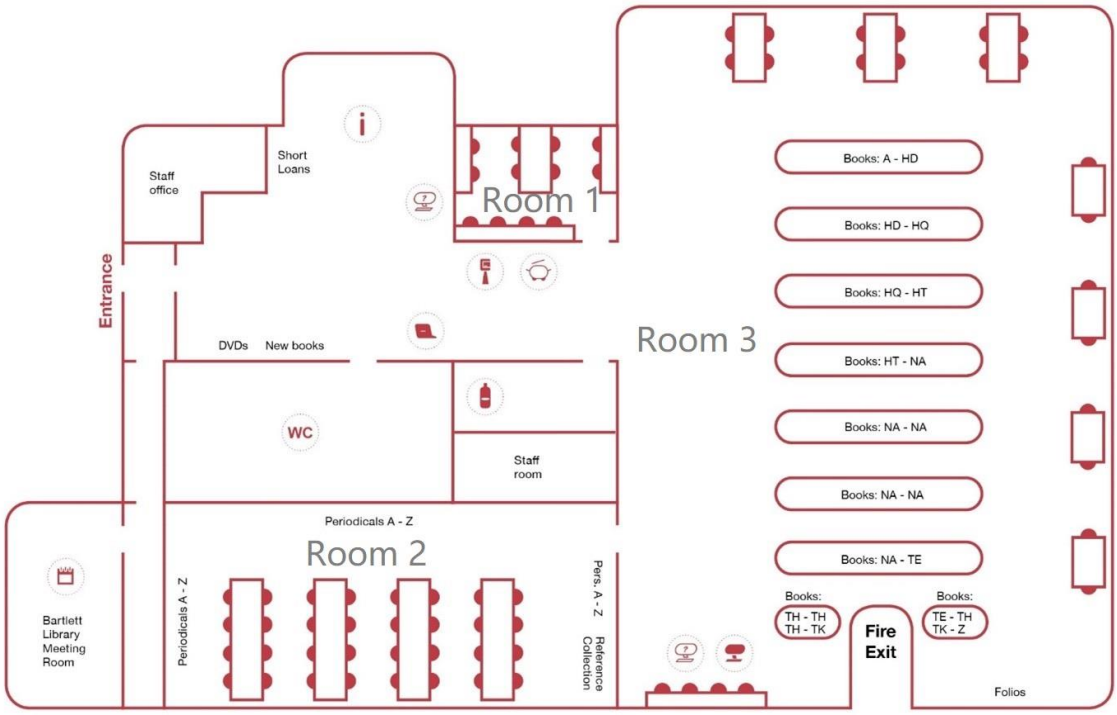


Figure 3 2D map of the Bartlett Library (UCL Bartlett Library Service, 2023)

3.3 Data Collection and Descriptive Statistics

3.3.1 Sensor Installation

In this study, I utilized sound sensors manufactured by BROWAN to measure and analyze sound pressure level (dBA) in various architectural settings. These sensors use LoRa® technology, a wireless modulation technology optimised for long-range

communication, while its upper layer network protocol, LoRaWAN™, provides these sensors with standardised and secure connectivity that ensures stable communication over long distances with low power consumption and support for a wide range of device deployments. LoRa® and LoRaWAN™ ensure that sensors can efficiently transmit their measurement data over long distances or in challenging environments. These sensors can monitor sound level ranging from 40 to 100 dB(A) in real-time, also featuring temperature detection capabilities from 0 to 50°C, with updates every 2 minutes. Detailed specifications of the sensor can be found in the Appendix B.

To comprehensively capture sound data in the Bartlett Library, we strategically placed 5 sensors in three primary areas of the library. This was to ensure thorough coverage of our main research zones. Each sensor was securely mounted at a roughly equivalent height to prevent potential damage and avoid any disturbance with the library users. In Room 1, the sensor was positioned on a protruding section of the wall, while in Rooms 2 and 3, they were placed at the top of bookshelves. The precise locations of these sensors are illustrated in the following graph (Figure 4).

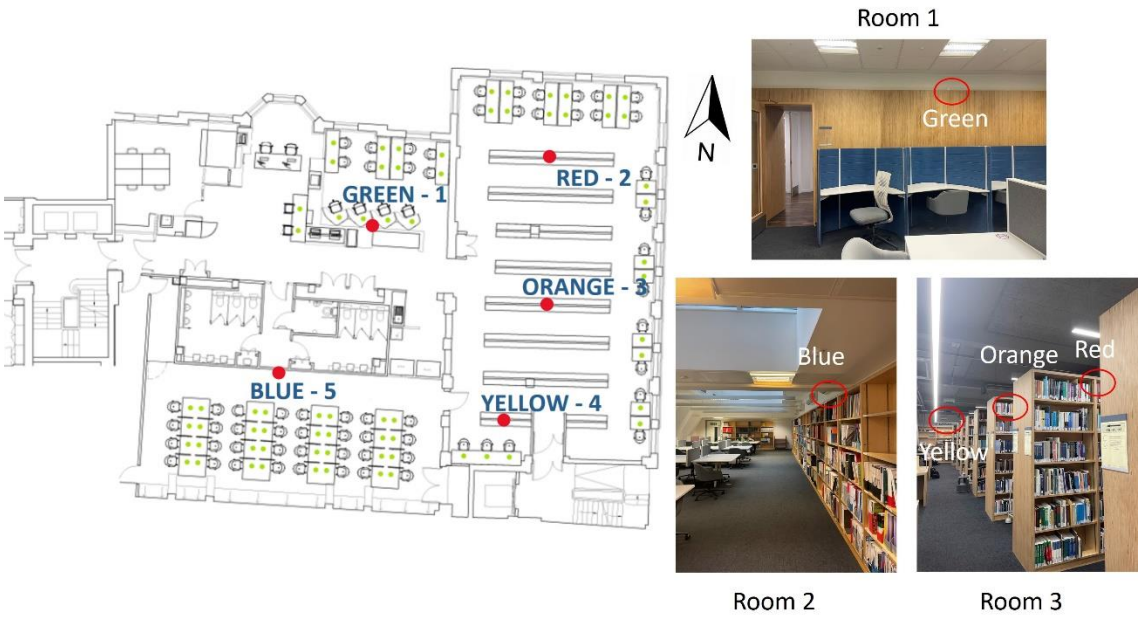


Figure 4 The Locations of Sound Sensor

Upon initial data review, I found that the Red sensor and Orange sensor, located in the Room 3, failed to collect sound data properly. As a result, data from these two sensors have been excluded from subsequent analyses.

When positioning these sensors, special attention must be paid to potential inaccuracies inherent to the sensors themselves. It's known that these sensors register a default level of 44 dBA. Based on validation, made with a professional Class 2 Isotech 52N Sound Level Meter, we have calculated the deviations from this default level, which are displayed in Table 3. While a laboratory-based calibration process for the various sensors was not feasible, due to cost and time constraints, the accuracy can still be deemed sufficient to meet the requirements of the conducted study. Consequently, before processing the sound level data, it is essential to adjust by

adding or subtracting the calculated deviation first. This ensures the accuracy of the data and eliminates potential discrepancies.

Table 3 Sensor Difference Parameters

	SENSOR ID	*NAME SENSOR*	*CALIBRATION VALUE*	*DIFFERENCE WITH REF VALUE 44* dBA
GREEN	eui-e8e1e1000104c51b	Tab 1	47dB(A)	+3
RED	eui-e8e1e1000104c5ca	Tab 2	43 dB(A)	-1
ORANGE	eui-e8e1e1000104c54b	Tab 3	48 dB(A)	+4
YELLOW	eui-e8e1e1000104c51e	Tab 4	50 dB(A)	+6
BLUE	eui-e8e1e1000104c538	Tab 5	48 dB(A)	+4

3.3.2 Noise Level Classification

The library, as a quiet area of public use, stands as an ideal place for fostering learning and concentration. In this study, I specifically measured and analyzed the noise environment within the library. My observations revealed that the primary noise sources in the library mainly include the sound of students discussing, street-facing ambient traffic noise, air conditioning noise and potential noise from facility renovations. To better understand the noise impact on the library environment, I conducted extensive measurements and analyses, tracking noise patterns over time. This step provides insights into the potential implications of noise in library. Given that noise level settings typically depend on specific use cases, I have established a set of evaluation criteria specifically designed for library environment of this case.

To measure the noise data, I utilized the sensor (Chapter 3.3.1 Sensor Installation) capable of recording sound level with a resolution of 1 decibel (dB). This device provides us with timely and accurate sound pressure level. It's worth noting that the data we collected are weighted, represented in units of dBA. The "A" indicates that the measurement has been adjusted to reflect the human ear's response to different frequencies. The dBA represents sound intensity at a specific moment. However, in environmental noise assessments, researchers typically focus on the L Aeq value, which is an average of the sound level over a period that considers energy accumulation. To transition from instantaneous sound measurements to L Aeq, employed the following conversion formula (Rossing, 2007):

First, calculate the squared sound pressure value for each instantaneous dBA value.

$$p_i^2 = 10^{(L_i/10)} \quad (3.1)$$

Next, compute the average squared sound pressure, where N is the number of measurement points.

$$\overline{p^2} = \frac{1}{N}\sum_{i=1}^{N} p_i^2 \quad (3.2)$$

Lastly, convert the average value.

$$L_{Aeq} = 10 \times \log_{10}(\overline{p^2}) \quad (3.3)$$

Regarding the noise level, I referenced the regulations from Building Bulletin 93 (Education Funding Agency, 2014) and the World Health Organization (Berglund et al., 1999), laying out the standards for this noise environment assessment shown in Table 4. The original regulations can be viewed in the Appendix C and D.

Table 4 Noise Level Classification

Level	*Duration (H)*	*Upper Limit [L Aeq (dBA)]*
1 Quiet	24	Less than 44
2 Moderate		45-74
3 Noticeable		75-84
4 Disturbing		Over 85

3.3.3 Spatial Attribution of Occupancy

The data collected in this study are all related to time series, but it is worth noting that each seat occupancy sensor has specific coordinate information on the UCL API, which means that we can use this data to explore the spatial distribution of seat occupancy, and even perform certain spatial analyses.

- Method for Getting Occupancy Data from UCL API

Firstly, log in to the UCL student account, after logging in we will get the exclusive key which can be used to access all the services in the UCL API Webpage[2]. The parameter setting codes can be seen from Figure 5.

```
survey_id: 8;
datetime__gte: 2023-06-02T00%3A00%3A00;
datetime__lte: 2023-07-30T00%3A00%3A00;
```

Figure 5[3] Parameter Setting Sample

The UCL API real-time data is updated every 20 minutes, while the historical data obtains the time and occupancy status each time the sensor switches state, therefore it is necessary to fill in the sensor data for the time when the state is not changed to achieve the same interval with the sound level data when conducting and sound level analysis in the later stage. Figure 6 is a sample of the sensor occupancy data obtained in JSON format.

[2] The data required for this study is obtained using the link:

https://uclapi.com/workspaces/historical/data&token=user-key

[3] Note: The survey_id is set to 8 which is the exclusive ID of the Bartlett Library, datetime__gte and datetime__lte are the start time and end time respectively, and T00%3A00%3A00 which is a time string in ISO 8601 format indicating 00:00:00.

```
{
    "20578001": {
        "okay": true,
        "data": {
            "next": null,
            "previous": null,
            "results": [
                {
                    "sensor_id": 20578001,
                    "datetime": "2023-06-02T00:00:00",
                    "state": 0
                },
            ]
        }
    }
}
```


Figure 6[4] Sample of Occupancy Original Data

Occupancy data in JSON format obtained in the previous step needs to use two libraries in Python, pandas and json, to flatten the data and then traverse the JSON data and then export the data in CSV format, and finally clean and standardise the data according to the theme requirements. After cleaning and standardising the occupancy data, the statistics are linked to the JSON data containing the geographic information data, then use geopandas, a library for processing and analysing geospatial data in Python, to call the shapely.geometry.Point class, which is used to represent two-dimensional point data, to extract the coordinate information and save it as a GeoJSON file, and eventually generate heat maps for each distribution topic using QGIS.

3.4 Time Series Analysis

Time series refers to a sequence of observations, x_t, arranged chronologically. A time series can be conceptualised as a set of sequentially indexed random variables

[4] Note: The state of the sensor (1=occupied; 0=absent; -1=unknown)

captured over time. For instance, a time series can be perceived as a set of random variables, x_1, x_2, x_3, ..., where x_1 denotes the value at the series' outset, and x_2 represents the value of the second point in time, continuing in this pattern (Brockwell and Davis, 2009). To ensure data integrity and accuracy, ideally, these observations should be equidistant and devoid of missing entries. Unlike other data sets, time series data introduces the unique dimension of time, which is both a constraint and a pivotal source of supplementary information.

Time series analysis is particularly adept at signal analysis. The sound level and seat occupancy data adopted in this study can be considered signals, making time series analysis beneficial in exploring the research question.

Time series analysis contains foundational exploratory data analysis and time series regression. When extending the regression analysis, predictive models can also be established. While exploratory data analysis primarily involves a rudimentary observation and analysis of the data, time series regression can be further utilised to build predictive models, forecasting future data trends. Subsequently, I will focus on the exploratory data study of time series.

3.4.1 Exploratory Data Analysis

Exploratory Data Analysis [5](EDA) encompasses the preliminary investigations into data to understand its fundamental characteristics, patterns, and anomalies. Before delving into EDA, it is essential to first undergo data cleansing.

● Data Processing

Sound Level Data:

1. Filter out data from the study period, which ranges from June 2nd, 2023, to July 30th, 2023.
2. Remove redundant columns. There are numerous outliers in room 3 from two sensors (Red and Orange).
3. Adjust the outlier values. Given that the sound sensor detection range is 40-100 dB(A), any value outside this range will be recorded as an outlier value of 255. Considering the slight possibility that the sound level in the library exceeds 100 dB(A), I adjusted the values to 255 to 39 dB(A).
4. Check for missing values and impute them.

Seat Occupancy Data:

[5] Note: The main aim of this study's time series analysis is to examine the relationship between sound level and seat occupancy data. Given their instantaneous nature, the analysis utilizes instantaneous sound levels, units using dB(A), rather than the previously referenced weighted average value L Aeq used to assess the overall noise level.

1. Filter out data from the study period. The first dataset is from June 2nd, 2023, to July 30th, 2023, and the second dataset is from June 18th, 2022, to July 9th, 2022.
2. Adjust the outlier values.
3. Check for missing values and impute them.
4. Link the room number with seat occupancy sensor IDs.
5. Compute the hourly occupancy count for each room. After learning the basics of the data, the next step is to perform stability and reliability tests.

● Augmented Dickey-Fuller Test

Augmented Dickey-Fuller (ADF) test, a statistical test known as the unit root test, which is designed to determine the extent to which a time series is affected by a trend. The original hypothesis (H0) of the ADF test assumes that the time series can be represented by a unit root and is not stationary. (Yamak, Yujian and Gadosey, 2019)

The results of the test for the sound level and seat occupancy statistics for each time period are shown in Table 5 below, based on the results, we can tell that they are all stationary and can be proceeded to the next step of the test. After testing the stationarity of the data, we need to test the autocorrelation of the data to determine if the data is completely random white noise.

Table 5 Results for Sound Level and Occupancy Statistics

Column	*ADF Statistic*	*p-value*	*Used lags*	*Number of observations used*	*1%*	*5%*	*10%*
OCCUPANCY 06/02-18	-3.56	0.03	23	1128	-3.97	-3.41	-3.13
VALUE 06/02-18	-3.65	0.03	20	1131	-3.97	-3.41	-3.13
OCCUPANCY 07/10-30	-3.80	0.02	23	1488	-3.96	-3.41	-3.13
VALUE 07/10-30	-12.79	2.56E-20	4	1507	-3.96	-3.41	-3.13

- White Noise Test

The goal of the white noise test is to determine whether a function time series is serially correlated in function space. In brief, it is to check if there is any pattern or predictability in the data or if it is just random noise (Zhang, 2016). The results of the test are shown in Table 6 below, from the results we can learn that all the variables p-values are extremely small, less than 0.5, which suggests that they are all autocorrelated, meaning that the data being tested is non-white noise data and not randomly generated. This also predicts that there may be some trend or periodicity in this data, so it is valuable to analyse it in the next step.

Table 6 Results for White Noise Test

Column	*Ljung-Box Statistic*	*P-value*
Occupancy 06/02-18	528.64	5.59E-117
Value 06/02-18	464.01	6.45E-103
Occupancy 07/10-30	731.60	4.01E-161
Value 07/10-30	223.07	1.94E-50

- Seasonal Decomposition

The nature of Seasonal Decomposition is to decompose the time series into trend, seasonality and noise components. This decomposition is particularly useful for detecting and characterising changes in the time series. The general model for seasonal decomposition is (Verbesselt et al., 2010):

$$Y_t = T_t + S_t + e_t \quad (3.4)$$

Where:

Y_t*: Observed data at time, t is time.*

T_t*: Trend component.*

S_t*: Seasonal component.*

e_t*: Remainder component (variation beyond the seasonal and trend components).*

After carrying out the exploratory data analysis, we can move on to explore the correlation between the two time series of Occupancy and Sound Level, and I will use several methods, linear correlation, non-linear correlation, and time series cross-correlation, to fully discuss the connection.

3.4.2 Correlation Analysis

- Pearson's Correlation Coefficient - Linear Correlation

Pearson's Correlation Coefficient assesses the linear relationship between two variables, ranging from negative linear relationship to positive linear relationship, which is from -1 to 1. For this study, it was applied to room classification (name: room), seat occupancy (name: Occupancy), and sound level to determine their linear correlations.

- Non-linear Correlation

In this study, the quadratic polynomial function is used instead to derive the non-linear relationship between occupancy and sound level.

- Cross Correlation - Time Series Correlation

Cross-correlation measures the similarity of two time series at different time lags, and its purpose is to test whether an event in one series affects the value of the other series in the future. This can help us identify potential lagged relationships between two time series. Cross-correlation results in a function or sequence (Welch, 1974). Its formula is:

$$R_\tau = \frac{\sum(x_t - \bar{x})(y_{t+\tau} - \bar{y})}{\sqrt{\sum(x_t - \bar{x})^2 \sum(y_t - \bar{y})^2}} \quad (3.5)$$

Where:

τ is time lags, x_t and y_t are data points from two time series.

$$R_\tau = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 \sum(y_i-\bar{y})^2}} \text{..(3.6)}$$

Where:

x_i and y_i are data points.

In fact, the formula for cross-correlation differs from the formula for Pearson's correlation coefficient only in, whether or not time is introduced.

3.5 Prediction Model

After conducting exploratory data analysis and correlation analysis, we discovered a nonlinear relationship between Occupancy and Sound Level (Analysis results are in Chapter 4.3 Is there any correlation between seat occupancy and sound level in the Bartlett Library?). Notably, these two variables exhibited evident trends and seasonality. Such pronounced dynamic temporal correlation indicates the necessity for a model adept at capturing this relationship for predictions.

While traditional time series models, such as the Autoregressive Moving Average (ARMA), Moving Average (MA) and Autoregressive Integrated Moving Average (ARIMA), perform well under certain circumstances, they are primarily designed for basic time series. Therefore, they struggle to capture and address complex time series relationships.

Considering the intricate relationship between Occupancy and Sound Level, we need a more advanced model capable of handling nonlinear time series data. Consequently, I opted for the Long Short-Term Memory (LSTM) and the Gated Recurrent Unit (GRU) models. Both of these models are extensions of Recurrent Neural Networks (RNN) and are particularly well-suited for challenges of these nature.

3.5.1 Long Short-Term Memory (LSTM) Model

Long Short-Term Memory (LSTM) is a special type of Recurrent Neural Network structure that is the standard structure used for machine learning analysis of time-series data, and is often used as a predictive model in the IoT domain due to its ability to recognise patterns over long sequences and its excellence in detecting anomalies, so it is adept at uncovering long-term dependencies (Das, ShuklaT and Sengupta, 2021).

And the ability of LSTM to handle long sequential data is due to the internal gate structure that determines what information needs to be kept in memory and what

information needs to be discarded. The basic LSTM design has three gates (input, forget, and output) and a cell state, making it capable of memorising or forgetting information over long periods of time. Due to its three-gate structure, the LSTM can control the flow of information more flexibly. the Input gate receives and processes data from the outside world on an on-going basis. Memory cell input gate takes its input from the output of the last iteration of the LSTM NN cell. The main role of the Forget gate is to control the time and lag, which allows it to choose when to forget the output result However, some researchers also add an input modulation gate, which exists to further modulate and filter the information deposited into the cell state (Fu, Zhang and Li, 2016), so that the LSTM model can be controlled in a more detailed way, and the specific LSTM cell-structure is shown in the following chart (Figure 7).

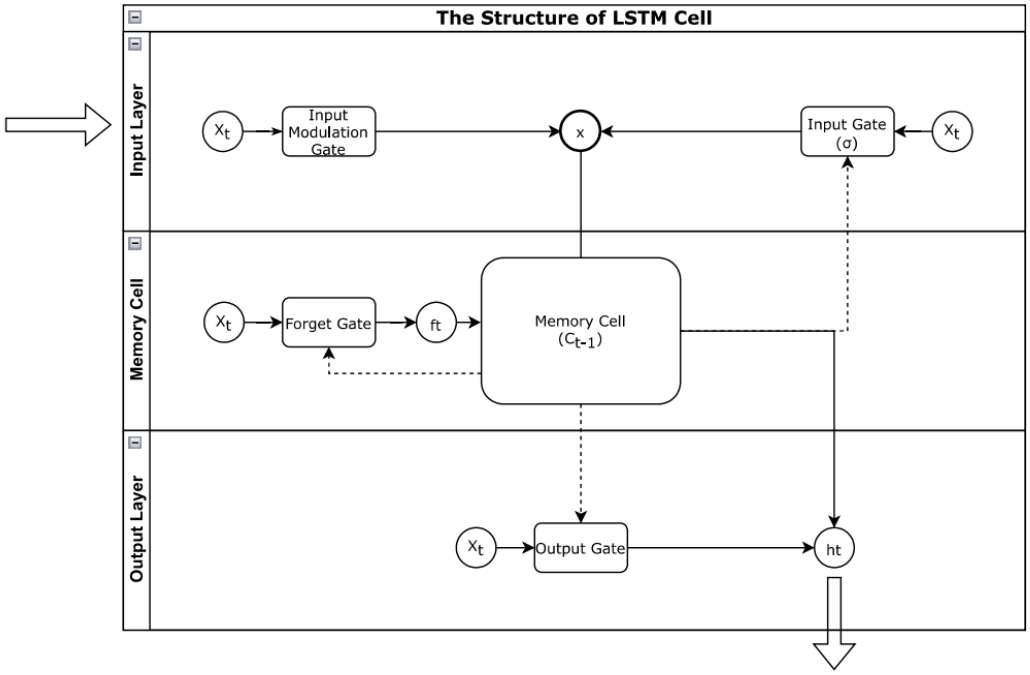


Figure 7 The Structure of LSTM Cell

Where:

h: points in hidden state of memory cell.

σ: represents the scalar product of two vectors or matrices.

3.5.2 Gated Recurrent Unit (GRU) Model

Gated Recurrent Unit (GRU) is a variant of recurrent neural network and also a simplified version of LSTM. Comparing to LSTM, GRU has one less gate, from 3 gates to 2 gates, namely Reset gate and Update gate. Update gate is more like a combination of Input gate and Forget gate in LSTM, which decides how much of the previous memories should be retained and how much of the new memories should be introduced. The Reset gate determines how much influence the previous memory should have at the current time step. These two gates ensure that the necessary information is retained in memory, and although the method has been streamlined, the accuracy and running speed have not been affected, and the computation is even faster than LSTM. The specific GRU cell structure is shown in Figure 8 below.

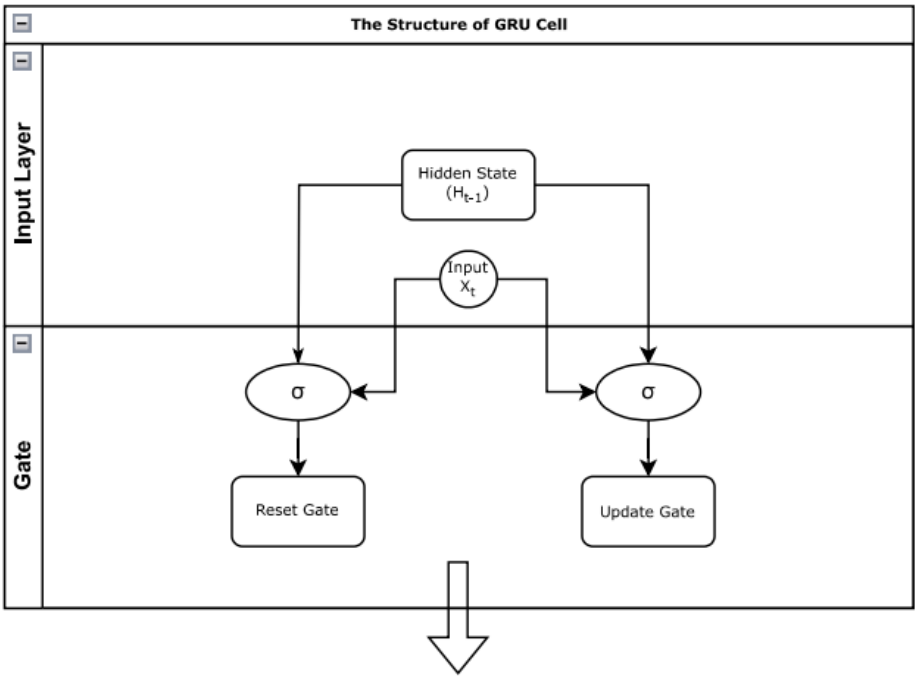


Figure 8 The Structure of GRU Cell

Summarising the above information, a comparison of the advantages and disadvantages of LSTM and GRU can be drawn as shown in the Table 7.

Table 7 Advantages and Disadvantages between LSTM and GRU

	Advantages	*Disadvantages*
LSTM	1. Flexible control of information flow. 2. Widely used and validated. 3. Expertise in handling long sequences.	1. Complexity of calculations. 2. Low training speed. 3. Too many parameters, prone to overfitting.
GRU	1. Simple structure of the model. 2. Few parameters and resources required for computation. 3. Fast training speed.	1. Not good at capturing long-term reliance. 2. Models are relatively new and have been applied and validated less frequently.

The prediction models of LSTM and GRU firstly use the sklearn library in Python to preprocess the combined data of occupancy and sound level, and then use the keras

library to construct the LSTM and GRU functions respectively, which set the independent variables to be the 'year', 'month', 'day', 'hour ', 'Value' (Sound level) columns in the data as well as the room column which is uniquely one-hot encoded. The dependent variable is the 'count_1s' column in the data, which is the occupancy. In addition to the basic variable settings, the following table is what parameters (initial values) need to be set (Table 8):

Table 8 Initial Parameter Setting

Parameter	*Model*	*Initial Value*	*Description*
look_back	LSTM/GRU	24	Consider the last 24 hours of data for prediction
Dropout	LSTM/GRU	0.2	A dropout rate of 20% is used to prevent overfitting
Epochs	LSTM/GRU	100	The model was trained 100 times on the entire dataset
Units	LSTM/GRU	50	50 neurons are used by the model in the LSTM and GRU layers
Batch size	GRU	32	32 samples of data were passed into the model at a time during training
Optimiser	LSTM/GRU	Adam	An optimisation algorithm with an adaptive learning rate

After training the model a prediction needs to be made, it should be noted that in case of predicting missing data, the occupancy data columns from the 18th of June to the 10th of July 2023 need to be removed in order to prevent affecting the model training.

At the end, the results are visualised using the matplotlib library (see Appendix E for main codes).

3.5.3 Validation

Because of the specificity of the data missing study, this paper sets up a total of three ways to validate the results:

1. Referring to the results of MSE, RMSE, MAE and R2 indicators.

2. Training the model and predicting the occupancy data of the data integrity time period (10 July 2023 to 30 July 2023) and comparing the predicted value with the true value. The comparison is done by calculating the absolute difference between the true value and the predicted value and then checking if the absolute difference is less than a preset threshold value (0.5), if it is less than or equal to, the result is returned as 'True', which is set to 0.5 because it is usually set to 0.5 in this type of study. Finally, the average of the true values is calculated. This mean value is used to assess the accuracy of the prediction results.

3. It was confirmed with the library staff that library usage does not fluctuate too much over the years, so the last method is to train the model and predict the occupancy data for the missing time period and compare it with the actual occupancy data from the previous year (2022).

The specific model selection still needs to be clarified by the next step of training and testing according to the specific scenario. Next chapter presents the results of the analyses and forecasting specific to this study.

3.6 Ethical Consideration

This study primarily revolves around the analysis of non-public data, none of which associates with personal information. The occupancy statistics of the Bartlett Library seats are derived from the UCL API, a third-party service for UCL members. It provides only the occupancy status of each seat within the library, without identifiable information. The sound level data was procured through sensors we installed, with prior permission from the Bartlett Library. These were strategically installed during low pedestrian flow to mitigate disruptions and were placed to not influence student activities. Importantly, these sensors, lacking audio recording functionality, pose no threats to privacy. The data collection process for this study was non-intrusive, causing no harm or inconvenience to library users. Moreover, the sensor data accrued, deemed research material, remain non-public data. Ethical clearance for this research was secured from University College London, approval number CASA23/5032999/1 (see in Appendix A).

4 Result and Analysis

Based on the methodology of (Chapter 3 Methodology), the results of the research question will be presented in this chapter.

4.1 How is the spatial distribution of seat occupancy and how is noise condition in the Bartlett Library during the research period?

4.1.1 Spatial Distribution of Seat Occupancy

This section is presented in the form of a heat map with four sub-figures, as shown in Figure 9. The first two maps compare weekday and non-workday seat occupancy during the study period. The next two maps show seat occupancy in the morning compared to the afternoon, where 2 p.m. is used as the dividing line. The reason for choosing 2 p.m. as the dividing line is to account for the fact that the Bartlett Library is open from 9 a.m. to 8 p.m. Using 2 p.m. as the dividing line ensures that the morning and afternoon time length are generally balanced, and that seat occupancy varies significantly from 2 p.m. onwards. In order to better compare the four subgraphs, I standardised the range of the data to 0 to 1. Because of the overall small size of the raw data values, directly using the raw values may result in an insignificant difference between the light and dark distributions of the heat maps, thus affecting the comparison. Therefore, a uniform standardisation to the 0 to 1 range allows for a clearer presentation and comparison of the data from each map.

The first set of heat maps shows that the frequency of seat occupancy is higher overall on weekdays than on Saturdays and clusters in the north window of room 3. On Saturdays, students are more likely to sit in the westernmost seat by the wall in room 2, and the seats in the middle part of room 1 and the third seat in the southeast corner of room 3 are more popular than on weekdays.

The second set of heat maps shows that overall seats are occupied more frequently in the morning (before 2 p.m.), and in the afternoon room 3 and room 1 become less occupied, but room 2 is occupied more frequently than in the morning.

Combining the two sets of maps, it can be seen that library users tend to sit in room 3, particularly the seats near the window, whereas room 2 is relatively less popular with users, and it can also be seen that the brightest seats in the maps are the ones outside room 1.

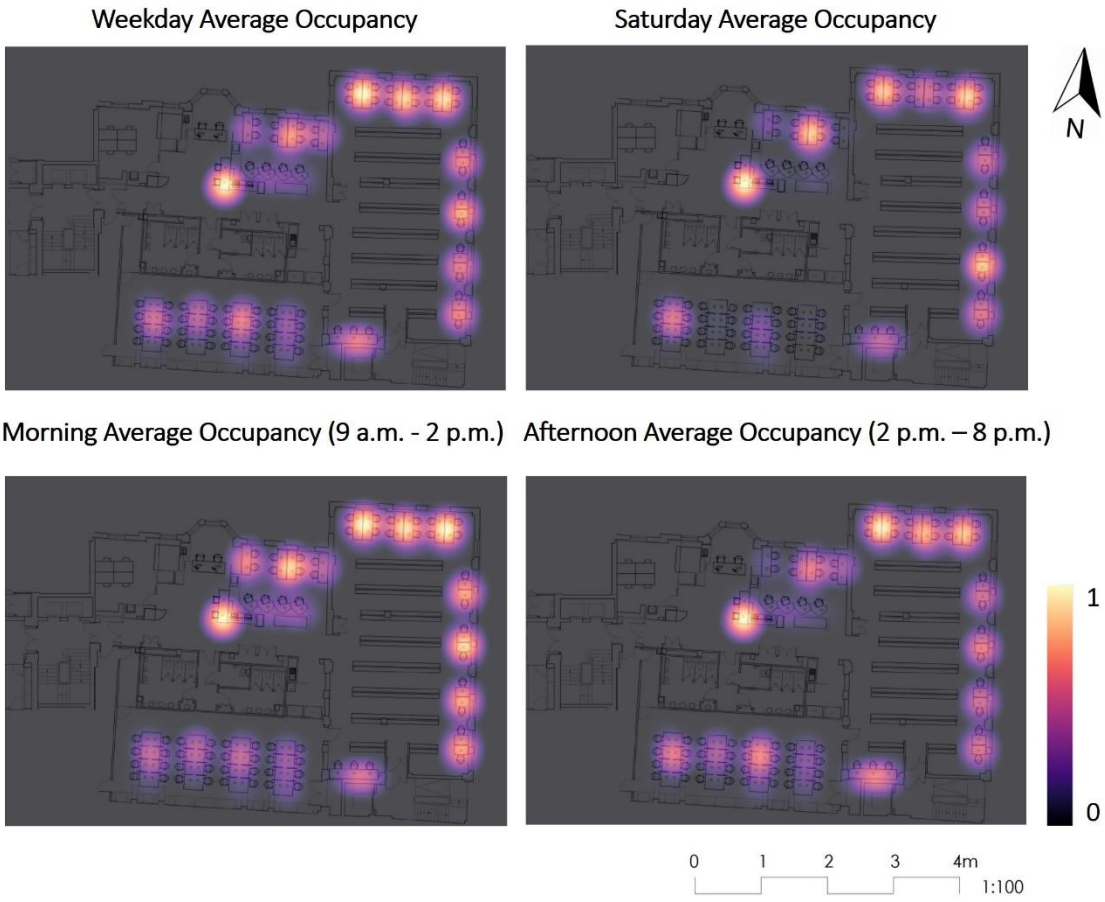


Figure 9 Heat map of the distribution of seat occupancy by different topics at the Bartlett Library

4.1.2 Noise Level Condition

The result of converting transient sound data to L Aeq using the method mentioned in (Chapter 3.3.2 Noise Level) is as follows. Given the cumulative nature of the data's energy, I opted for an area chart representation. As depicted in Figure 10, there's a notable daily average noise level variation in room 1 and room 2 throughout the observation period, while room 3 remains relatively stable. Room 1 experiences several noise peaks between June 17th to June 23rd and July 10th to July 21st, with June 19th spiking to approximately 93 dBA, falling into a disturbing noise level. The

peak for room 2 is on July 14th, reaching 92 dBA. In addition to this, I calculated the basic statistics for the noise level data, Table 9 shows that average noise level for room 1, room 2, and room 3 during the observation period are 63.90, 70.53, and 56.52 dBA respectively. It also can be inferred that room 1 has a noise data standard deviation of 26.14 and a variance of 684.29, the highest among three rooms, indicating the most significant noise fluctuations. Room 1 also has the highest number of days with noticeable and disturbing Noise Level, making its noise issue the most apparent. Room 2's noise fluctuations days are at a moderate level. It occasionally gets disturbing, but overall, it is more stable than room 1. Room 3 is the most stable and quietest, with minimal noise level fluctuations and very few days with noticeable or disturbing noise level. In terms of noise, room 3 might be the most comfortable room.

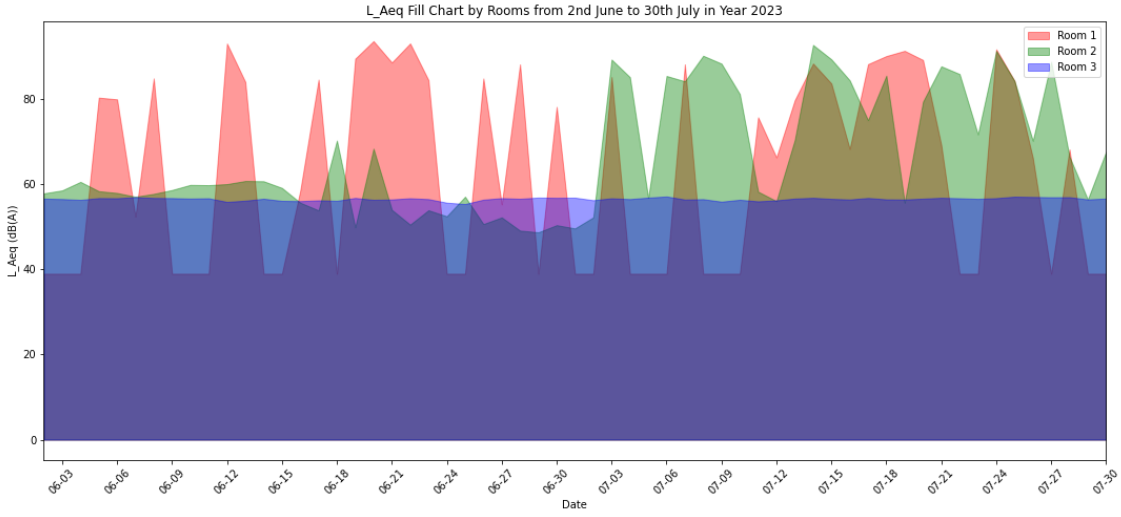


Figure 10 Noise Level Fill Chart for Different Rooms

Table 9 Results for Noise Level Statistics

	S.D.	*Variance*	*Maximum*	*Minimum*	*Average*	*25th*	*50th*	*75th*	*Noticeable noise level days*	*Disturbing noise level days*
Room 1	26.16	684.29	93.66	39.00	63.90	39.00	56.92	85.20	8	12
Room 2	11.14	124.07	92.76	48.73	70.53	56.82	57.97	60.80	6	5
Room 3	0.39	0.15	57.15	55.38	56.52	56.44	56.67	56.83	0	0

Note: S.D. is Standard Deviation.; 25 th, 50 th, 75 th mean 25 th Percentile, 50 th Percentile, 75 th Percentile

4.2. What are the temporal trends of sound level and seat occupancy in the Bartlett Library during the study period?

4.2.1 Time Series Results

Firstly, as Figure 11 and Figure 12 shown below, there are time series charts for the two data collection periods. Segmenting the data into three sub-plots based on rooms, we can intuitively observe the fluctuations in sound level and seat occupancy over time.

During the first period, the graph for room 1 clearly shows three major fluctuations in sound level, peaking at close to 100 dB(A) on June 12th. It is worth noting that some of the apparent occupancy changes coincide with the sound level fluctuations and peak on June 6th at around 10 occupancies. The sound level data for room 2 remains stable, except for a noticeable drop on June 9th. However, we can observe a similar

drop in the sound data of room 3 on the same day. Occupancy is lower in room 2 and peaks on June 13th at about 7 occupancies. In contrast, in room 3, occupancy is outstanding overall, peaking at about 13 occupancies in a single hour. However, in room 2 and 3, the trend of occupancy and sound level fluctuations do not match as well as in room 1. In addition, from the charts, we can clearly identify non-workdays, such as June 4th and June 11th, in which the seat occupancy is zero.

In the second research period, the sound data for room 1 displays frequent fluctuations from July 12th to July 21st, followed by scattered changes after July 24th. The sound level data for room 2 exhibits larger fluctuations overall but has relatively smaller changes between July 11th and July 13th. In contrast, room 3's sound level remains relatively stable, around 55 dB(A). Compared to the data from June, with the exception of room 3, the sound level in the other two rooms increase and demonstrate pronounced variability. As for seat occupancy, the data for July shows a noticeable increase compared to June, with room 3 experiencing the most frequent changes in occupancy, with up to 15 occupancies in a single hour. Of the three rooms, only room 1's sound level and occupancy fluctuations match each other. We can also identify non-workdays during this period, such as July 17th, July 24th and July 30th, where the seat occupancy is registered as zero.

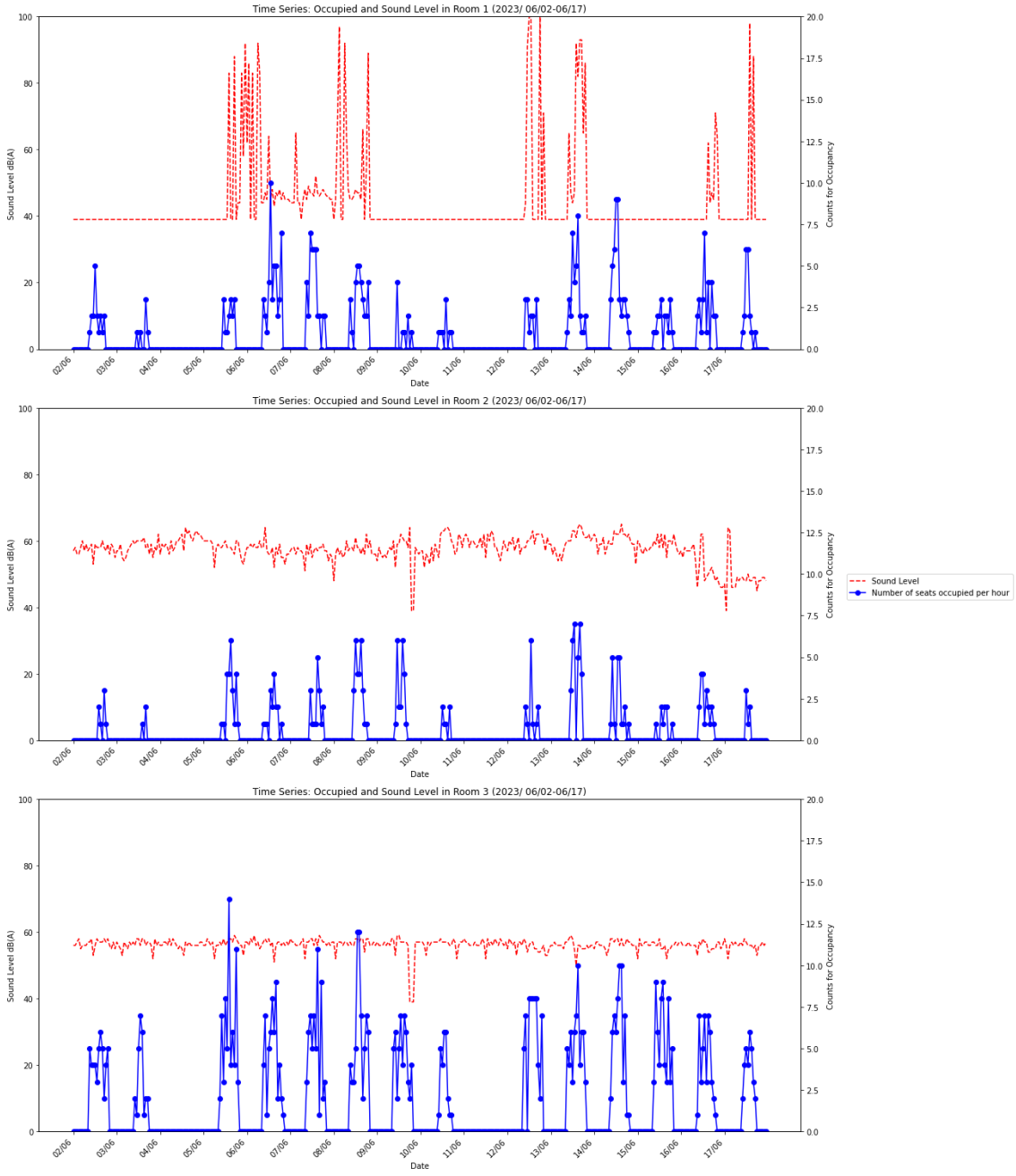


Figure 11 Time Series for Sound Level and Occupancy in Different Rooms (2023/ 06/02-06/17)

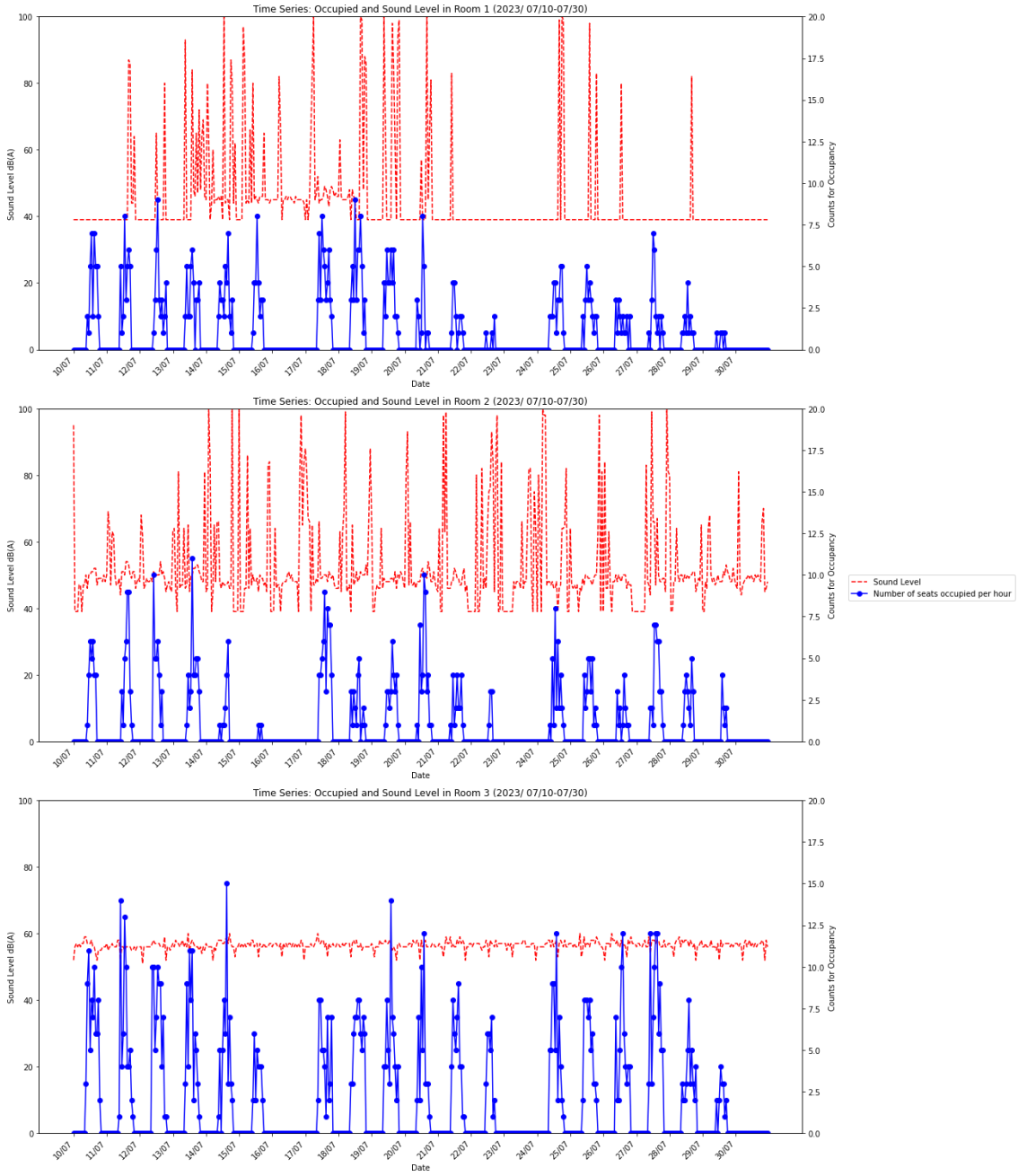


Figure 12 Time Series for Sound Level and Occupancy in Different Rooms (2023/ 07/10-07/30)

Apart from that, I also calculated the descriptive statistics as shown in Table 10 below, this result helps us to better understand the overall situation of the two data sets. And from this table, we can draw the following conclusions:

1. Seat occupancy in July is more frequent than in June, yet the overall environment is quieter.
2. The variance increases from 4.71 in June to 6.07 in July, and the standard deviation rose from 2.17 in June to 2.46 in July. This suggests that the data for July is relatively more dispersed.
3. Covariance can be used to measure how two sets of data change together. The two covariance values in the table are 2.54 and 3.06, both positive, which means that when occupancy increases in this time series, the sound level also tends to increase, and vice versa, which means that they are to some extent changing in the same direction.

Table 10 Results for Sound Level and Occupancy Statistics

	Mean		*Median*		*Mode*		*S.D.*		*Covariance*	
Time Period	June	July	June	July	June	July	June	July	June	July
Occupancy	1.11	1.31	0.00	0.00	0.00	0.00	2.17	2.46	2.54	3.06
Sound Level	52.50	51.13	56.00	49.00	39.00	39.00	9.94	12.02		

Note: June is 2023 06/02-18 and July is 2023 07/10-30; covariance is for Occupancy and Sound Level

4.2.2 Seasonal Decomposition Results

From Figure 13 and Figure 14, we can see that neither the sound level data nor the occupancy data for both periods show a clear and consistent upward or downward trend, except that they rise on weekdays and then fall again towards the weekend. The seasonality is also quite uniform, both being on a one-day cycle, peaking at about 1 pm. In the first period, the residuals of the occupancy of room 3 and the sound level of room 1 and in the second period, the residuals of the occupancy of room 3 and the

sound level of room 1 and room 2 fluctuated considerably, but these fluctuations do not present a clear pattern and the means are close to 0. Overall, the seasonal decomposition succeeds in extracting the main seasonal effects and the trend effects from the original plots.

Figure 13 Seasonal Decomposition for Sound Level and Occupancy (2023/ 06/02-06/17)

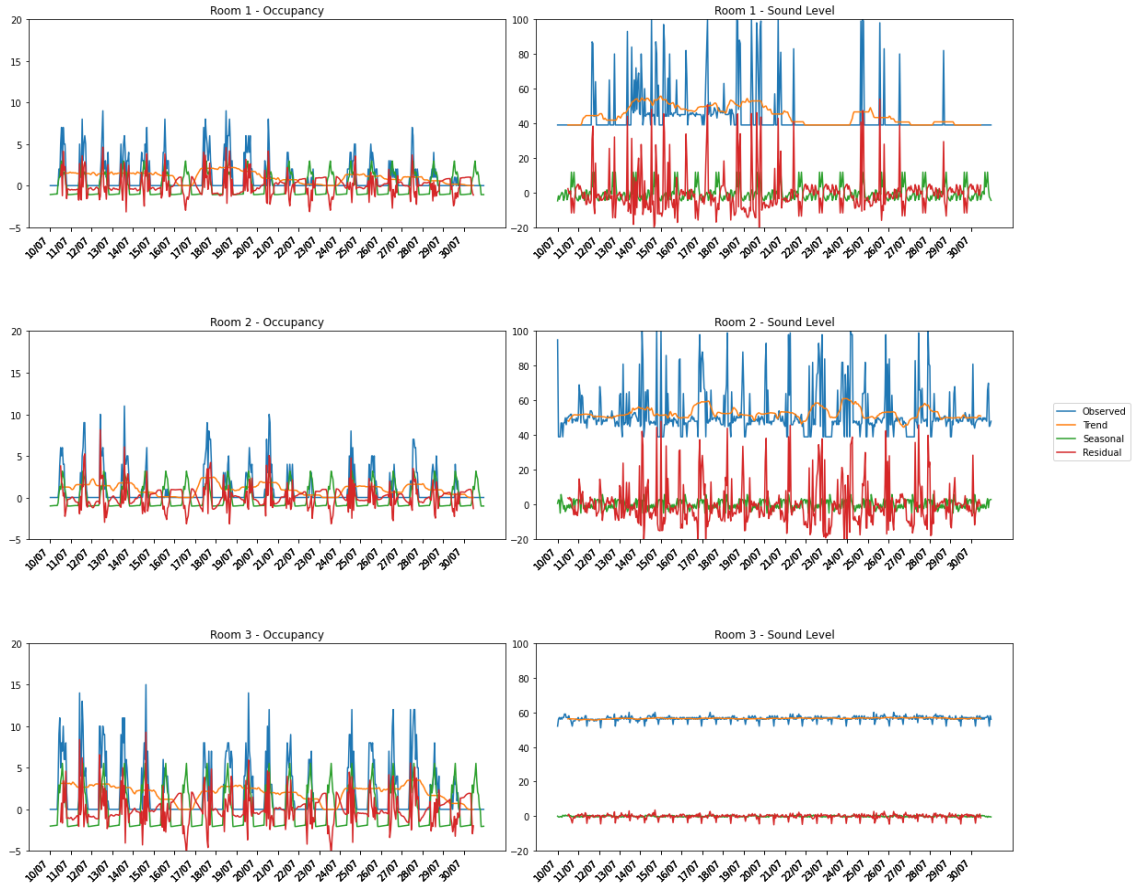


Figure 14 Seasonal Decomposition for Sound Level and Occupancy (2023/ 07/10-07/30)

4.3 Is there any correlation between seat occupancy and sound level in the Bartlett Library?

4.3.1 Linear Correlation

The first step was to use Pearson's coefficient to check the linear relationship between the variables and the results were obtained as figure shown below (Figure 15).

From the Pearson correlation coefficients across two periods, the relationship between room and sound level is most evident, with coefficients of 0.51 and 0.41. This suggests some rooms might be noisier than others. While the strength of the linear correlation varies, both datasets show a positive correlation. Notably, there's no apparent linear relationship between occupancy and sound level, which prompts further exploration into their non-linear relationship.

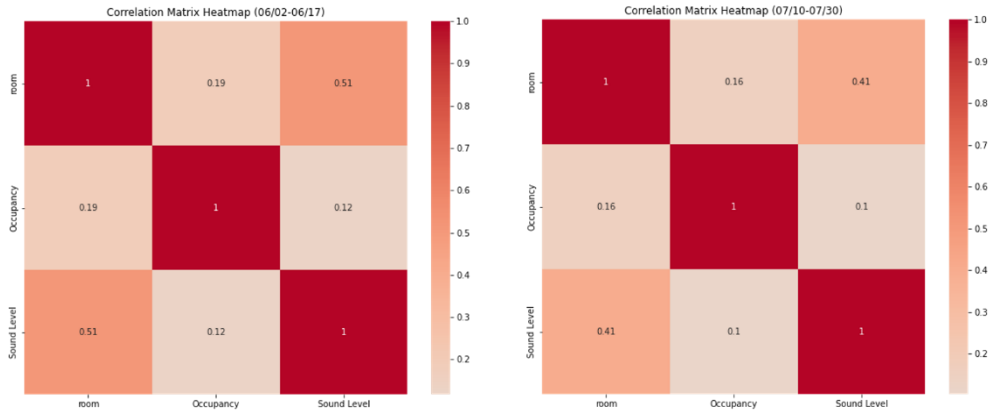


Figure 15 Pearson's Correlation Coefficient Heatmap

(06/02-17 on the left; 07/10-30 on the right)

4.3.2 Non-linear Correlation

As can be seen from Figure 16, the resulting curves are curved upwards, which shows that sound level increases as occupancy increases, but slows down as occupancy becomes larger. However, there are still some points in the graphs that are not successfully fitted, and these may be affected by other factors.

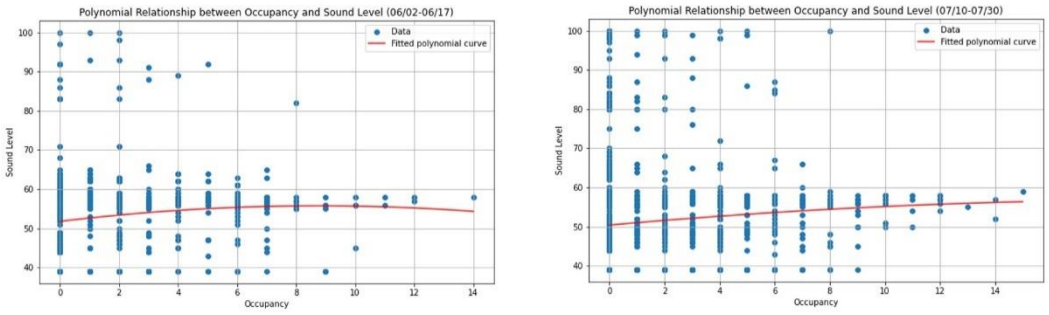


Figure 16 Polynomial Relationship between Occupancy and Sound Level

(06/02-17 on the left; 07/10-30 on the right)

4.3.3 Time Series Correlation

After determining the linear and non-linear relationship, with the knowledge from (Chapter 3.4.1 Exploratory Data Analysis) that the two sets of time series are stationary and are not noise data. In this case, it is possible to go on to determine the similarity of these time series under different time lags. The result is obtained as shown in Figure 17, where the X-axis represents the delays (lags), which represents the time difference between the two sequences. The Y-axis represents the correlation.

In the first period, the data initially oscillate around 0.1, indicating a weak positive correlation between occupancy and sound level. However, as x exceeds 400, this relationship becomes unstable and moves towards a negative correlation until it reaches a low of -0.6 at x = 850. The bounce at x = 900 may be influenced by external factors.

The second period of the data set oscillates around 0, showing that there is no clear linear relationship between the two series, but the increase in the amplitude of the oscillations implies an increase in the instability of the data.

Overall, the first dataset demonstrates a complex relationship between occupancy and sound level, whereas the second dataset is more stable but may be influenced by external factors.

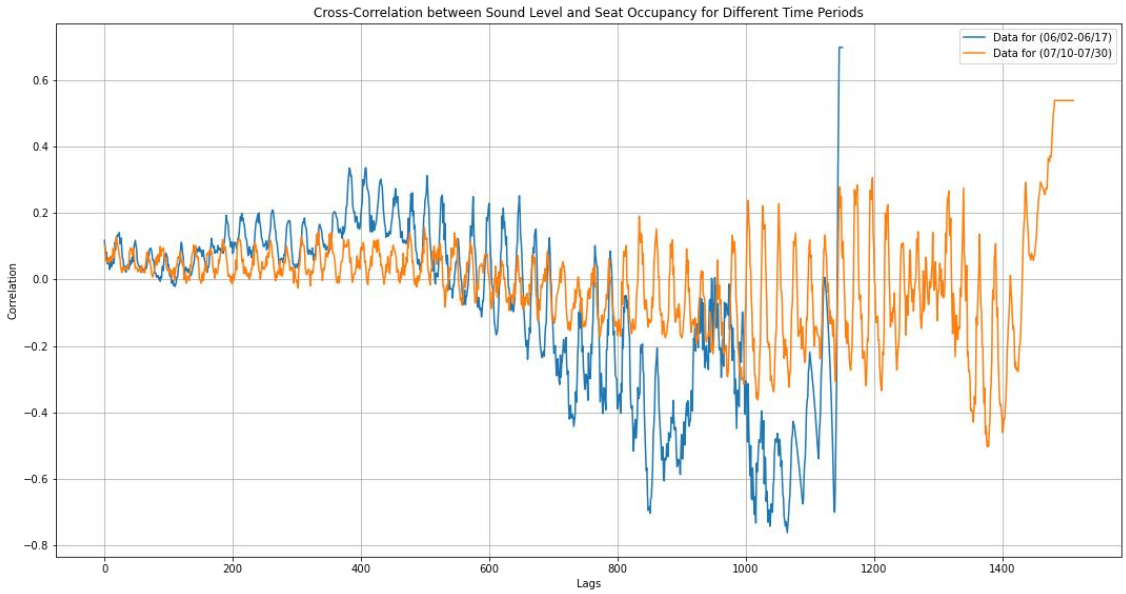


Figure 17 Cross-Correlation between Sound Level and Seat Occupancy for Different Time Periods

4.4 How to impute the missing occupancy data for the Bartlett Library with the assistance of the relationship between occupancy and sound level?

We can observe an oscillatory pattern in the resultant (Figure 17) of the cross-correlation analysis. The periodicity occurs approximately 8 times out of 200 time

points of data, which indicates that there are approximately 25 time points in each cycle. Combining the opening hours and usage patterns of the library, and taking into account the periodicity of the data obtained from the previous analysis, 24 time points can be identified as the optimal lag period. In time series forecasting, choosing the right lag period (Referring to the explanation of parameter settings for the LSTM and GRU models in Chapter 3.5 Prediction Model) is critical because it determines how much past data we are going to use to predict future values. Considering that the unit of this experiment is the hour, a 24-hour look-back parameter is selected for the LSTM and GRU models. The cross-validation results of the two models are as follows:

In cross-validation, both Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) measure the magnitude of the prediction error, with smaller values indicating better model performance. Mean Absolute Error (MAE) represents the average level of error. The closer the R-squared (R2) is to 1 the more accurate the prediction is. And the training time for this cross-validation of LSTM and GRU models is 30 minutes and 20 minutes respectively. As can be seen from the results in Table 11, GRU outperforms LSTM in all metrics except MAE and is faster to train, resulting in GRU model being the better choice.

Table 11 Cross-validation Result for LSTM and GRU

	LSTM				GRU			
	MSE	*RMSE*	*MAE*	*R2*	*MSE*	*RMSE*	*MAE*	*R2*
Fold 1	2.0575	1.4344	0.7953	0.2170	1.8236	1.3504	0.7144	0.3060
Fold 2	4.8260	2.1968	1.2502	0.2194	4.7508	2.1796	1.2975	0.2316
Fold 3	2.4362	1.5608	0.8695	0.4359	2.3489	1.5326	0.8429	0.4561
Fold 4	3.3973	1.8432	0.9118	0.0940	3.2411	1.8003	0.9131	0.1357
Fold 5	5.2374	2.2885	1.2184	0.4760	5.1586	2.2712	1.4606	0.4839

After multiple iterations, the optimized GRU model adopted the following hyperparameters: 'look back' of 24, indicating it requires 24 time steps for predictions; 'units' of 50, ensuring the model is neither too simple nor overfitted; and a 'dropout' value of 0.4 to mitigate overfitting by randomly deactivating 40% of the neurons during training. The optimiser chooses the "Adam". Using these settings and the prediction accuracy comparison method mentioned in (Chapter 3.5.3 Validation), I obtained the results presented in the table below (Table 12).

The R2 values on different folds fluctuate from 0.1113 to 0.4914, which indicates that this model has better predictive performance in some subsets. The maximum values of MAE and MSE are from fold 5, it can be learnt that fold 5 may have a special pattern or noises that cause the model's prediction performance to be degraded. The average value of accuracy within threshold is 0.4488, which means that about 44.88% of all predictions have an error between the predicted value and the true value within the given threshold value. In general, the GRU model shows some stability across different data subsets, but there is still scope for improvement.

Table 12 Cross-validation and Accuracy Result for GRU (After Changing Hyperparameter)

	MSE	*RMSE*	*MAE*	*R2*	*Accuracy within Threshold*
Fold 1	2.0243	1.4228	0.7226	0.2296	0.4964
Fold 2	3.7261	1.9303	1.1090	0.3973	0.3915
Fold 3	2.3499	1.5329	0.8352	0.4559	0.4641
Fold 4	3.3326	1.8255	0.9139	0.1113	0.5251
Fold 5	5.0836	2.2547	1.2974	0.4914	0.3668
Mean	3.3033	1.7932	0.9756	0.3371	0.4488
S. D.	1.2139	0.3308	0.2285	0.1613	0.0677

The model is known to present some stability, then I took the occupancy and sound level data from 10th July to 30th July as the true values and ran the model again to predict the occupancy values for that time period and compared the true values with the predicted values and got the results as in Figure 18.

From Figure 18, it can be seen that the predicted value is more average and all lower than the true value, but the predicted value shows obvious periodicity, which is 24 hours as a cycle, but for non-working days of the data prediction accuracy is not considerable, Sunday's occupancy value should be 0, but the predicted value is not 0.

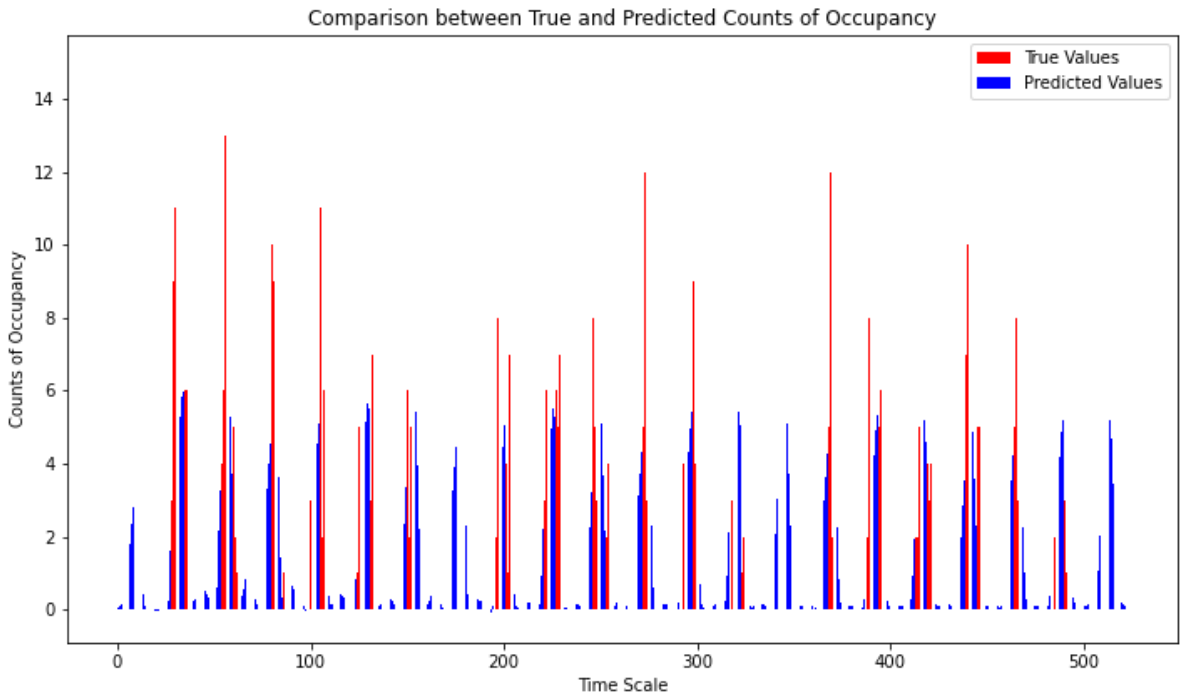


Figure 18 Comparison between True Values and Predicted Value of Occupancy in Year 2023

Given that this is already the best model that could be trained with the currently available data, I apply it to estimate seat occupancy for the period 18th June to 9th July 2023, a period for which raw seat occupancy data were missing and could not be directly validated. In order to assess the predictive effectiveness of the model, I sought other validation methods. After consulting with library staff, I learned that seat occupancy rates change relatively steadily from summer to summer. Therefore, I chose the seat occupancy data from the same period last year as a reference benchmark to compare with the model's predictions and obtained the results shown in Figure 19.

As can be seen from the result (Figure 19), the first third of the data predicted is more in line with the real value, the difference in the middle part is large, and the last part of

the data successfully predicted an upward trend compared to the first two paragraphs, but there is still a significant gap with the actual value.

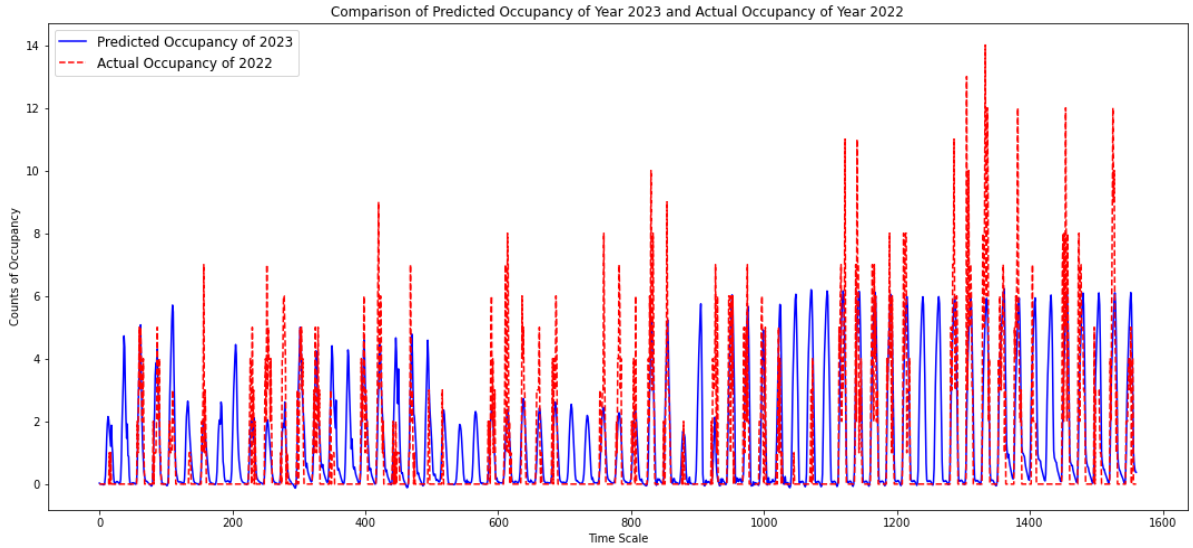


Figure 19 Comparison of Predicted Value of Year 2023 and True Value

5 Discussion

5.1 Review of results

The purpose of this study is to explore the spatial distribution of seat occupancy and sound level in Bartlett Library over the study period, as well as the temporal trends of both. In addition, I am concerned with whether there is a correlation between seat occupancy and sound level in the Bartlett Library and explore how this relationship can be used to help impute missing seat occupancy data for the library. The results in Chapter 4 indicate the following:

5.1.1 Seat Occupancy and Noise Condition in Library Rooms

Room 1: While this room is also popular in terms of seat occupancy, its noise level varies widely and often reaches disturbing level.

Room 2: Neither the most popular nor the quietest, its seat occupancy lies between room 1 and room 3, and noise level occasionally reaches disturbing level in July.

Room 3: This room is the most popular and quietest in the entire library. The noise level is stable on moderate level. This could mean that users prefer a quiet environment for studying.

However, I found some peculiarities in Figure 9 and Figure 11 in Chapter 4 that need to be explained: For instance, one of the seats outside of room 1 is unusually busy,

which is later confirmed to be an office area, which explains why it is used so frequently. The sudden drop in sound level in Figure 11, on the other hand, could be due to a temporary malfunction of the sensor or the user being too far away from the sensor and thus failing to capture the sound.

It is worth noting that even on the 19th of June when sound level fluctuated greatly in room 1 and room 2 due to a fire alarm (verified by asking staff), the sound level in room 3 remained steady. Upon further examination, I determine that the sound sensor in room 3 may not have been accurately capturing sound data due to its isolated location in the room (southwest corner) or the sensor's own sensitivity.

5.1.2 Observations on Time Trends

Overall, seat occupancy was more frequent in July than in June, but sound level was lower than in June. Both months' data show an increase in weekdays versus a decrease in weekends, and seat occupancy and sound level show a degree of synchronisation between weekdays and non-workdays, which means it is possible to distinguish between working and non-working days by those two datasets. Seasonality and trends have been successfully extracted from the data, and there is no clear overall upward or downward trend.

5.1.3 Relationship Analysis

Whilst in some cases a positive correlation is shown between seat occupancy and sound level, this relationship can be destabilised by external factors or time periods. The relationship between them is not simply linear, but rather non-linear, as evidenced by an exponential growth relationship. However, there are still some data points that do not fit well on the exponential curve, which may be affected by complex seasonal factors.

5.1.4 Estimating Missing Data Using Deep Learning Models

I attempt to estimate the missing data using the GRU model. The prediction results are fair, but there is some deviation from the actual values. Despite my efforts in data preprocessing to eliminate outliers and ensure data smoothing, the possibility of overfitting due to insufficient data size cannot be ruled out. More accurate prediction results may be obtained if data with longer time ranges are available.

5.2 Significance

According to the previous literature review (Chapter 2 Literature Review), Tunahan and Altamirano also explored the spatial occupied patterns of the Bartlett Library and compared the changes before and after the COVID-19 pandemic, which is shown in Figure 20. The results of my study are highly consistent with their occupancy patterns of the spaces, especially the occupancy patterns before the epidemic, which proves

the accuracy of my analysis of occupancy, and indirectly indicates that the epidemic has reduced the impact of the library usage patterns.

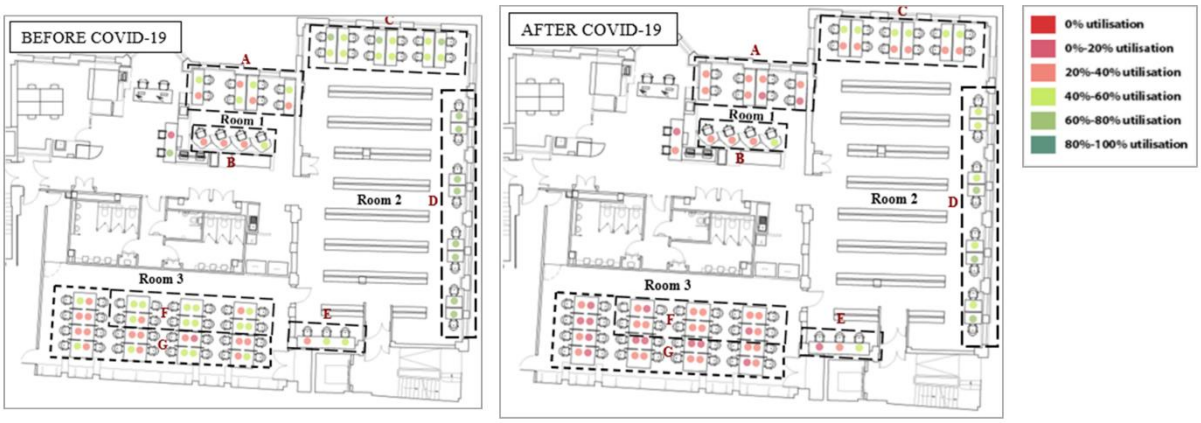


Figure 20 Change in the utilisation of the desks at the UCL Bartlett Library before and after the COVID-19 pandemic. (Tunahan and Altamirano, 2022)

Differing from previous studies, this study also adds a new variable, sound level, to analyse the relationship between occupancy and sound level in depth. In contrast to the unidirectional studies by Ekwevugbe's team (Ekwevugbe, Brown and Fan, 2012; Ekwevugbe et al., 2013), I found a positive correlation and complex time-series correlation between sound and occupancy, which influenced each other, this finding not only validates the possibility of estimating occupancy using sound data, but also provides a basis for subsequent estimation of occupancy data using deep learning.

I further develop a predictive model to estimate missing occupancy data using sound level time series data. While there are still shortcomings in this approach, the use of environmental data such as sound level to fill in the missing data for occupancy is

indeed a novel and worthwhile direction to explore in depth. This approach ensures the completeness and reliability of the study and provides a richer data basis for subsequent studies.

5.3 Limitations and the Future

In the process of this study, there are still some unavoidable or not easily solved limitations, although the best efforts were made to avoid them.

5.3.1 Observation period limitation

The observation period of this study is only 59 days, and considering the missing data situation, the actual effective observation is only 37 days. Such a time period is relatively short for a study that mainly relies on time series analysis. The short observation period may lead to a more complex correlation between the two sets of time series data, as time series models have a limited ability to capture patterns on short-term data. Future studies should extend the observation period to obtain more representative and robust data.

5.3.2 Model Limitations

This study of predictive data used a more sophisticated deep learning model, but it is an undeformed original model. According to a previous literature review, deep learning models can be better adapted to training and test data by combining them with transfer

learning (Chen et al., 2021). Future research could consider integrating this approach to optimise the model for more accurate predictions.

5.3.3 Hardware constraints

This study is constrained by the sensor hardware, as some of the installed sensors fail to work properly and suffer from insensitive sensor detection. To ensure the accuracy and completeness of the data, future studies should consider using more stable and sensitive sensors. In addition, selecting sensors that can collect multiple environmental data will provide researchers with more comprehensive information, which will not only help to verify the operational status of the sensors in real time, but also further explore the connection between multiple environmental factors and space.

6 Conclusion

On a broader level, the purpose of this paper is to explore trends in occupancy and sound level over time within the Bartlett Library and the relationship between them, on the basis of which missing occupancy data can be predicted.

First, in order to carefully assess the spatial patterns of library use, this study utilises heat maps to depict the distribution of seat occupancy in the Bartlett Library. In addition, fill maps and statistical analyses illuminate the noise condition of the library. The results tell us that: Room 1, despite being popular, frequently experiences disturbing noise level; Room 2 has a moderate level of seat occupancy and occasionally reaches disturbing noise level in the month of July, whereas Room 3 is the quietest and most popular. Overall, library users seem to prefer a quieter study environment. And occupancy patterns differ between weekday and non-weekday, morning and afternoon.

Next, time series plots and seasonal decompositions are used to reveal temporal trends in sound level and seat occupancy. Notably, neither shows a consistent long-term trend. This study delves into the relationship between occupancy and sound level using tools such as Pearson's coefficient and polynomial functions. There is a non-linear relationship between seat occupancy and sound level, which is complex in its presentation over time and is influenced by external factors.

Finally, based on the relationship derived above, the time series prediction models LSTM and GRU models are used to impute the missing seat occupancy data in Bartlett Library, and the performance of the models is evaluated using methods such as cross-validation, and comparison of predicted data with real data. The validation results illustrate that compared to the LSTM model, the GRU model performs better in terms of prediction and shows some stability on certain subsets of data, but there is still scope for improvement of the GRU model, and there are some gaps compared to the real values.

From the point of view of the application of IoT technology, the complete research ideas and methods of this paper, taking the Bartlett Library as an example, obtains the data in a non-invasive way and explains the connection between occupancy and sound level. These in-depth studies provide library users and managers with a comprehensive understanding of Bartlett Library. For users, this information can help them choose a more appropriate place to study quickly, while for managers, this data not only enhances their knowledge of the library, but also contributes to more efficient management and optimisation. From an interior building research perspective, the analytical and predictive methods used in the paper are transferable and can be used in similar indoor studies, which provides ideas and examples for other researchers. Looking to the future, refined measurements could enhance future studies. This may require the deployment of more robust sensors with a longer observation duration. In addition to this, the use of methods such as transfer learning to aid model optimisation may be beneficial in obtaining more reliable missing data in the future.

7 Reference

Ang, I.B.A., Dilys Salim, F. and Hamilton, M. (2016) 'Human occupancy recognition with multivariate ambient sensors', in *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*. *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, Sydney, Australia: IEEE, pp. 1–6. Available at: https://doi.org/10.1109/PERCOMW.2016.7457116.

Berglund, B., Lindvall, T. and Schwela, D.H. (1999) Guidelines for Community Noise, World Health Organization. Available at: https://www.who.int/publications/i/item/a68672 (Accessed: 22 August 2023).

Brockwell, P.J. and Davis, R.A. (2009) *Time Series: Theory and Methods*. Springer New York (Springer Series in Statistics). Available at: https://books.google.co.uk/books?id=TVIpBgAAQBAJ.

Camastra, F. *et al.* (2022) 'Prediction of environmental missing data time series by Support Vector Machine Regression and Correlation Dimension estimation', *Environmental Modelling & Software*, 150, p. 105343. Available at: https://doi.org/10.1016/j.envsoft.2022.105343.

Cha, S.H. and Kim, T.W. (2015) 'What Matters for Students' Use of Physical Library Space?', *The Journal of Academic Librarianship*, 41(3), pp. 274–279. Available at: https://doi.org/10.1016/j.acalib.2015.03.014.

Chen, Z. *et al.* (2021) 'A transfer Learning-Based LSTM strategy for imputing Large-Scale consecutive missing data and its application in a water quality prediction system', *Journal of Hydrology*, 602, p. 126573. Available at: https://doi.org/10.1016/j.jhydrol.2021.126573.

Das, T., ShuklaT, R.M. and Sengupta, S. (2021) 'Imposters Among Us: A Supervised Learning Approach to Anomaly Detection in IoT Sensor Data', in *2021 IEEE 7th World Forum on Internet of Things (WF-IoT)*. *2021 IEEE 7th World Forum on Internet of Things (WF-IoT)*, New Orleans, LA, USA: IEEE, pp. 818–823. Available at: https://doi.org/10.1109/WF-IoT51360.2021.9595280.

Deb, C. *et al.* (2017) 'A review on time series forecasting techniques for building energy consumption', *Renewable and Sustainable Energy Reviews*, 74, pp. 902–924. Available at: https://doi.org/10.1016/j.rser.2017.02.085.

Denzin, N.K. and Lincoln, Y.S. (2005) The sage handbook of qualitative research. Thousand Oaks: Sage Publications.

Department for Environment, F.& R.A. (2010) Noise policy statement for England, GOV.UK. Available at: https://www.gov.uk/government/publications/noise-policy-statement-for-england (Accessed: 23 August 2023).

Dong, B. *et al.* (2019) 'A review of smart building sensing system for better indoor environment control', *Energy and Buildings*, 199, pp. 29–46. Available at: https://doi.org/10.1016/j.enbuild.2019.06.025.

Education Funding Agency (2014) BB93: Acoustic design of schools - performance standards, GOV.UK. Available at: https://www.gov.uk/government/publications/bb93-acoustic-design-of-schools-performance-standards (Accessed: 22 August 2023).

Ekwevugbe, T. *et al.* (2013) 'Real-time building occupancy sensing using neural-network based sensor network', in *2013 7th IEEE International Conference on Digital Ecosystems and Technologies (DEST)*. *2013 7th IEEE International Conference on Digital Ecosystems and Technologies (DEST) - Complex Environment Engineering*, Menlo Park, CA, USA: IEEE, pp. 114–119. Available at: https://doi.org/10.1109/DEST.2013.6611339.

Ekwevugbe, T., Brown, N. and Fan, D. (2012) 'A design model for building occupancy detection using sensor fusion', in *2012 6th IEEE International Conference on Digital Ecosystems and Technologies (DEST)*. *2012 6th IEEE International Conference on Digital Ecosystems and Technologies (DEST) - Complex Environment Engineering*, Campione d'Italia, Italy: IEEE, pp. 1–6. Available at: https://doi.org/10.1109/DEST.2012.6227924.

Fisk, W.J. (2000) 'H EALTH AND P RODUCTIVITY G AINS FROM B ETTER I NDOOR E NVIRONMENTS AND THEIR R ELATIONSHIP WITH B UILDING E NERGY E FFICIENCY', *Annual Review of Energy and the Environment*, 25(1), pp. 537–566. Available at: https://doi.org/10.1146/annurev.energy.25.1.537.

Frontczak, M. and Wargocki, P. (2011) 'Literature survey on how different factors influence human comfort in indoor environments', *Building and Environment*, 46(4), pp. 922–937. Available at: https://doi.org/10.1016/j.buildenv.2010.10.021.

Fu, R., Zhang, Z. and Li, L. (2016) 'Using LSTM and GRU neural network methods for traffic flow prediction', in *2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC)*. *2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, Wuhan, Hubei Province, China: IEEE, pp. 324–328. Available at: https://doi.org/10.1109/YAC.2016.7804912.

Fu, T. (2011) 'A review on time series data mining', *Engineering Applications of Artificial Intelligence*, 24(1), pp. 164–181. Available at: https://doi.org/10.1016/j.engappai.2010.09.007.

Gilani, S. and O'Brien, W. (2017) 'Review of current methods, opportunities, and challenges for in-situ monitoring to support occupant modelling in office spaces', *Journal of Building Performance Simulation*, 10(5–6), pp. 444–470. Available at: https://doi.org/10.1080/19401493.2016.1255258.

Gao, Y. (2023) GaoYuanru/Casa_dissertation, GitHub. Available at: https://github.com/GaoYuanru/CASA_Dissertation.git (Accessed: 24 August 2023).

Gou, Z., Khoshbakht, M. and Mahdoudi, B. (2018) 'The Impact of Outdoor Views on Students' Seat Preference in Learning Environments', *Buildings*, 8(8), p. 96. Available at: https://doi.org/10.3390/buildings8080096.

Hodgson, M. and Nosal, E.-M. (2002) 'Effect of noise and occupancy on optimal reverberation times for speech intelligibility in classrooms', *The Journal of the Acoustical Society of America*, 111(2), pp. 931–939. Available at: https://doi.org/10.1121/1.1428264.

Jadhav, D. and Shenoy, D. (2020) 'Measuring the smartness of a library', *Library & Information Science Research*, 42(3), p. 101036. Available at: https://doi.org/10.1016/j.lisr.2020.101036.

Keskin, Z., Chen, Y. and Fotios, S. (2015) 'Daylight and seating preference in open-plan library spaces', The International Journal of Sustainable Lighting, 1(1), p. 12. doi:10.17069/ijsl.2015.12.1.1.12.

Kim, K. (2015) 'Sources, Effects, and Control of Noise in Indoor/Outdoor Living Environments', *Journal of the Ergonomics Society of Korea*, 34(3), pp. 265–278. Available at: https://doi.org/10.5143/JESK.2015.34.3.265.

Liang, X. (2018) 'Internet of Things and its applications in libraries: a literature review', *Library Hi Tech*, 38(1), pp. 67–77. Available at: https://doi.org/10.1108/LHT-01-2018-0014.

Liu, J. *et al.* (2020) 'Human Occupancy Detection via Passive Cognitive Radio', *Sensors*, 20(15), p. 4248. Available at: https://doi.org/10.3390/s20154248.

Izmir Tunahan, G., Altamirano, H. and Unwin Teji, J. (2021) The impact of daylight availability on seat selection, UCL Discovery - UCL Discovery. Available at: https://discovery.ucl.ac.uk/id/eprint/10137716/ (Accessed: 25 August 2023).

Ma, J. *et al.* (2020) 'A bi-directional missing data imputation scheme based on LSTM and transfer learning for building energy data', *Energy and Buildings*, 216, p. 109941. Available at: https://doi.org/10.1016/j.enbuild.2020.109941.

Madakam, S., Ramaswamy, R. and Tripathi, S. (2015) 'Internet of Things (IoT): A Literature Review', *Journal of Computer and Communications*, 03(05), pp. 164–173. Available at: https://doi.org/10.4236/jcc.2015.35021.

Meyn, S. *et al.* (2009) 'A sensor-utility-network method for estimation of occupancy in buildings', in *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*. *2009 Joint 48th IEEE Conference on Decision and Control (CDC) and 28th Chinese Control Conference (CCC)*, Shanghai, China: IEEE, pp. 1494–1500. Available at: https://doi.org/10.1109/CDC.2009.5400442.

Minoli, D., Sohraby, K. and Occhiogrosso, B. (2017) 'IoT Considerations, Requirements, and Architectures for Smart Buildings—Energy Optimization and Next-Generation Building Management Systems', *IEEE Internet of Things Journal*, 4(1), pp. 269–283. Available at: https://doi.org/10.1109/JIOT.2017.2647881.

Pedersen, T.H., Nielsen, K.U. and Petersen, S. (2017) 'Method for room occupancy detection based on trajectory of indoor climate sensor data', *Building and Environment*, 115, pp. 147–156. Available at: https://doi.org/10.1016/j.buildenv.2017.01.023.

Petersen, S. *et al.* (2016) 'Establishing an image-based ground truth for validation of sensor data-based room occupancy detection', *Energy and Buildings*, 130, pp. 787–793. Available at: https://doi.org/10.1016/j.enbuild.2016.09.009.

Pourahmadi, M. (1989) 'ESTIMATION AND INTERPOLATION OF MISSING VALUES OF A STATIONARY TIME SERIES', *Journal of Time Series Analysis*, 10(2), pp. 149–169. Available at: https://doi.org/10.1111/j.1467-9892.1989.tb00021.x.

Rossing, T. (2007) *Springer Handbook of Acoustics*. Springer New York (Springer Handbook of Acoustics). Available at: https://books.google.co.uk/books?id=4ktVwGe_dSMC.

Sahu, V. and Gurjar, B.R. (2021) 'Spatio-temporal variations of indoor air quality in a university library', *International Journal of Environmental Health Research*, 31(5), pp. 475–490. Available at: https://doi.org/10.1080/09603123.2019.1668916.

Thomas, G. (2011) 'A Typology for the Case Study in Social Science Following a Review of Definition, Discourse, and Structure', *Qualitative Inquiry*, 17(6), pp. 511–521. Available at: https://doi.org/10.1177/1077800411409884.

Tunahan, G.I. and Altamirano, H. (2022) 'Seating Behaviour of Students before and after the COVID-19 Pandemic: Findings from Occupancy Monitoring with PIR Sensors at the UCL Bartlett Library', *International Journal of Environmental Research and Public Health*, 19(20), p. 13255. Available at: https://doi.org/10.3390/ijerph192013255.

UCL Bartlett Library , L.S. (2023) UCL Bartlett Library Floor Plan, Login. Available at: https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fwww.ucl.ac.uk%2Flibrary%2Fsites%2Flibrary%2Ffiles%2Fbartlett-library-floor-plan.docx&wdOrigin=BROWSELINK (Accessed: 22 August 2023).

Verbesselt, J. *et al.* (2010) 'Detecting trend and seasonal changes in satellite image time series', *Remote Sensing of Environment*, 114(1), pp. 106–115. Available at: https://doi.org/10.1016/j.rse.2009.08.014.

Wang, Q., Patel, H. and Shao, L. (2023) 'A longitudinal study of the occupancy patterns of a university library building using thermal imaging analysis', *Intelligent Buildings International*, 15(2), pp. 62–77. Available at: https://doi.org/10.1080/17508975.2022.2147129.

Wang, W., Chen, J. and Hong, T. (2018) 'Occupancy prediction through machine learning and data fusion of environmental sensing and Wi-Fi sensing in buildings', *Automation in Construction*, 94, pp. 233–243. Available at: https://doi.org/10.1016/j.autcon.2018.07.007.

Wei, Y. *et al.* (2022) 'Indoor occupancy estimation from carbon dioxide concentration using parameter estimation algorithms', *Building Services Engineering Research and Technology*, 43(4), pp. 419–438. Available at: https://doi.org/10.1177/01436244211060903.

Welch, L. (1974) ‘Lower bounds on the maximum cross correlation of signals (Corresp.)’, *IEEE Transactions on Information Theory*, 20(3), pp. 397–399. Available at: https://doi.org/10.1109/TIT.1974.1055219.

‘Wilson Committee Report on Noise’ (1963) *Aircraft Engineering and Aerospace Technology*, 35(8), pp. 218–219. Available at: https://doi.org/10.1108/eb033763.

Xia, J. (2005) ‘Visualizing occupancy of library study space with GIS maps’, *New Library World*, 106(5/6), pp. 219–233. Available at: https://doi.org/10.1108/03074800510595832.

Yamak, P.T., Yujian, L. and Gadosey, P.K. (2019) ‘A Comparison between ARIMA, LSTM, and GRU for Time Series Forecasting’, in *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*. *ACAI 2019: 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, Sanya China: ACM, pp. 49–55. Available at: https://doi.org/10.1145/3377713.3377722.

Zhang, X. (2016) ‘White noise testing and model diagnostic checking for functional time series’, *Journal of Econometrics*, 194(1), pp. 76–95. Available at: https://doi.org/10.1016/j.jeconom.2016.04.004.

Zhou, Y. *et al.* (2020) ‘A novel model based on multi-grained cascade forests with wavelet denoising for indoor occupancy estimation’, *Building and Environment*, 167, p. 106461. Available at: https://doi.org/10.1016/j.buildenv.2019.106461.

Zotoo, I.K., Lu, Z. and Anyigbah, G.L.E. (2023) 'The Impact of Library Seats and Noise on Students’ Perception of the Library Environment: A Study of Jiangsu University Library', Current Journal of Library and Information Sciences (CJLIS), 1(1).

8 Appendix

Appendix A Form B: Low Risk Ethics Application & Data Protection Registration

The content of the attachment below is the original and approved Ethics Application Form, Approval number CASA23/5032999/1.

CASA Ethics Procedure for MSc/MRes Dissertations

Form B: Low Risk Ethics Application & Data Protection Registration

What this form is about

You should complete this form only if *Form A: Screening of Ethical Risk* identified your research as low risk; if you have not yet completed it, please do not proceed and complete Form A first.

The purpose of this form is to

- Record details on the low ethical risk of your research, including – where applicable – risks to both potential participants and yourself;
- Specify how your research abides to the principles of benefit and (no) harm, informed consent and confidentiality, as described in *Form A: Screening of Ethical Risk*;
- Specify ways in which you responsibly address ethical risks in your research, including – where applicable – a protocol for managing pre-collected data, Personal Data and Sensitive Data.

If your research is low risk, you must submit this form and receive an approval notice before you can proceed with your research.

How to complete this form

This form contains seven sections covering:

1. General information about your project
2. Research methods
3. Location, Permissions & Risks to Researchers
4. Details of participants
5. Accessing and processing pre-collected data
6. Processing Personal Data and Sensitive Data
7. Declaration

Please note that sections 4, 5 and 6 may not be applicable to your research. Depending on your research methods, you might need to develop and provide a Participant Information and Consent Sheet. There will be clear instructions in the form as to which sections you need to complete and what additional information you need to provide.

Please note, if your research involves the recruitment of participants, you should not offer any monetary compensation or incentives other than confirming to share a copy of your final research report or providing coffee or refreshments during the interview.

If you have any questions while filling in the form, please contact your supervisor.

1. General information

Student name:	Student email address:	Student number:
Yuanru Gao	ucfnaoh@ucl.ac.uk	22227790
Supervisor name:	**Supervisor email address:**	**Date:**
Valerio Signorelli	ucfnvsi@ucl.ac.uk	26/04/2023

Dissertation research project title or topic:

How external environmental factors affect seating selection of students: Exploring from PIR Sensors data at the UCL Bartlett Library

2. Research methods

2.1. Provide a brief background (including aims) to the project in plain English (300 words max).

This project aims to investigate how changes in sensory factors, such as noise, affect students' use of study spaces in the library. The study utilizes data on seat vacancies collected by the UCL Bartlett Library PIR sensor, provided by UCL API, and noise data detected by sensors installed inside the library (to be installed) and detectors on the roof of the building. The research begins with a spatial analysis of the indoor seat availability data to identify the spatial distribution and preferences of students in the library. The noise data is then analyzed to identify any fluctuations in sound levels during the library's opening hours. Finally, the relationship between noise and seat occupancy is examined, including how the occupancy rate changes in response to noise and how it is distributed throughout the library. The research aims to develop an automated approach for assessing students' selection of study spaces.

2.2. Which methods are you going to use?

Tick all that apply.

- [x] Secondary data analysis (pre-collected data)
- [x] Collection/use of sensor or locational data
- [ ] Interviews
- [ ] Focus groups
- [ ] Questionnaires (including surveys and oral questions)
- [ ] Observation of participants in their own environment
- [ ] Observation of participants in a different environment
- [ ] Action research, with the observer participating
- [ ] Documentary analysis (systematic analysis of written material, this can include the use of personal records)
- [ ] Audio/visual recordings (including photographs)
- [ ] Controlled Trial
- [ ] Intervention study (including changing environments)
- [ ] Systematic review (of published research)
- [ ] Advisory/consultation groups
- [ ] Other, please give details:

2.3. Provide an overview in plain English of the project (500 words max).

Please focus on your research design and describe what data or samples you will collect, how you will collect data, what topics you will cover and what you will ask any participants to do. You should justify your chosen methods.

The key focus of this project is to investigate the impact of noise on the distribution of students' study locations. The required data include the UCL API-provided library vacancy information, the Bartlett Library rooftop noise detector, and the Bartlett Library internal noise detector (pending permission and installation). The UCL API data is available for UCL students to develop programs, and can be accessed and analyzed directly using tokens and links. The external noise detector is a Rion NL-52 Long-Term Noise Measurements device that outputs data in CSV format. Datasets will be made available on a monthly basis, named after the site where the station is located, and include noise indicators such as LAeq, LAFMax, LAFMin, LCeq-LAeq, LAleq-LAeq, LA1, LA10, LA50, LA90, and one-third octave bands at 1-minute resolution. Frequency analysis and trend of changes in noise levels will be conducted based on the noise data, and the vacancy data will be combined with the noise data over time for further analysis.

3. Location, permissions & risks to researchers

→ *UCL has a duty of care to all its staff and students under the Health and Safety at Work Act*

3.1. Will you conduct all or part of your research outside the UK?

- ○ **YES.** Go to the next question.
- ◉ **NO.** Go to question 3.3.

3.2. If the research includes work outside the UK, is ethical approval in the host country (local ethical approval) required?

See Guidelines for Research Conducted Overseas.

- ○ **YES.** Please specify the countries and confirm whether ethical approval has been received. If applicable, attached a copy of local ethical approval.
- ○ **NO.** Please confirm if local ethical approval will be sought or why it is not necessary.

3.3. In what type of location will your research take place?

For example, in public spaces, offices, private properties, hotels, via video call etc.

UCL premises, UCL Bartlett Library

3.4. Is permission required to conduct your research at the locations you list above?

(●) **YES.** Please explain how this permission will be obtained prior to data collection.

Email to UCL Bartlett Library and inquire whether the library accepts the installation of new sensors and what information is required.

( ) **NO.** Go to the next question.

3.5. During the research, will you conduct the research in any of these places?

	YES	NO
Alone in a non-public place (e.g. dwellings, workplaces with very few workers present)?	( )	(●)
Alone in a public place with few other people present (e.g. quiet park/street)?	( )	(●)
In a place where the research topic might be considered sensitive?	( )	(●)
Overseas in an area where the UK Foreign, Commonwealth & Development Office (FCDO) advises against travel (amber / red on the FCDO map of that country).	( )	(●)

*If you answered **YES** to any of the question above, you must complete a Risk Assessment.*

☐ Confirm that you will complete a risk assessment. Contact CASA's Teaching Team (casa-teaching@ucl.ac.uk) and attach the completed risk assessment to this application.

4. Details of participants

→ *Participants may only be included in your research if they have given informed consent to participate.*

4.1. Are you planning to recruit participants into the study?

( ) **YES.** Describe how potential participants will be recruited.

This should include reference to how you will identify and approach participants. For example, will participants self-identify by responding to an advert for the study or will you approach them directly, such as in person or via email?

***Now**: Use this template to develop a process to inform participants about your study and seek their consent to participate. Provide a brief description of the process here:*

(●) **NO.** Go to section 5.

4.2. How will you record consent?

Tick all that apply

- ☐ Noting consent at the start of *an Anonymous Verbal Survey*
- ☐ Recording consent at the start of *a Written Questionnaire or Survey* [highly recommended]
- ☐ Recording consent to interviews or focus groups using Email [highly recommended]
- ☐ Recording consent *Verbally* [highly recommended]
- ☐ Recording consent using a *Signed or Initialled Form*
- ☐ Other, please describe:

4.3. Please state any *benefits* to participants in taking part in the study

This may include feedback, access to services or incentives.

4.4. Please state any *risks* to participants and how these risks will be addressed.

5. Accessing and processing pre-collected data

→ *Unless available in the public domain, pre-collected data may only be used if permissions have been obtained from data owners.*

5.1. Does your study involve the use of previously collected data?

- ◉ **YES.** Please list the names of all data sources that you are intending to use and specify their owners (where available you may wish to provide a URL for describing/accessing the data).

  1. UCL API :https://uclapi.com/
  2. Sound Levels Monitoring Project in London (UK) : https://rdr.ucl.ac.uk/collections/Sound_Levels_Monitoring_Project_in_London_UK_/5307266

- ○ **NO.** Go to section 6.

5.2. Are all the datasets in the public domain?

Material that any member of the public is (legitimately) free to access and use, without having to obtain permission from anyone else, would be considered as being in the public domain.

- ○ **YES.** Go to section 6.
- ◉ **NO.** Go to the next question.

5.3. Do you have the owners' permission to use their data?

- ◉ **YES.** Please specify how the permission was given (e.g. through a data sharing agreement) and attach the agreement to this form, if applicable.

  UCL API's authentication (for both developers & end users) is done via the UCL login system. So I can access to the data by using my student information.

- ○ **NO.** Please explain how you intend to obtain permission or why permission is not needed.

5.4. Will you be conducting analysis within the remit the data were originally collected for?

- ◉ **YES.**
- ○ **NO.** Please explain how consent was gained from participants for further analysis and describe how you address any ethical issues that may arise from your use of data outside their remit.

6. Personal Data and Sensitive Data

→ *Adhere to fair processing principles when using Personal Data or Sensitive Data*

6.1. Are the data you intend to use fully anonymised?

- ○ **NO,** it may be possible to identify individuals (including indirectly/by chance). ⇒ Go to question 6.2.
- ○ **NO,** but it will not be possible to identify individuals. ⇒ Go to question 6.4.
- ◉ **YES.** ⇒ Go to question 6.4.

6.2. What types of Personal Data are you collecting and processing?

*If you collect and process data that are not anonymised or that may allow identification of individuals indirectly or by chance, then you are processing **Personal Data**. Personal Data comprise data pertaining to a person who could directly or indirectly be identified from that data, including data that you are collecting simply to contact your participants, such as names, residential addresses, email addresses, telephone number or IP addresses. You may only collect or process Personal Data if absolutely necessary and if appropriate safeguards are in place.*

Important: If you intend to process Personal Data, please complete the following training: UCL Data Protection for Researchers and Students.

Please list the Personal Data items that you are planning to collect and process:

If your research has been registered for Data Protection elsewhere, please state the institution and the registration reference code:

6.3. Please confirm that you will conscientiously adhere to the following fair processing practices.

(a) You will process the data fairly and lawfully abiding by UCL Data protection guiding principles.

(b) You will collect and/or use the minimum personal data necessary for the research.

(c) You will only use the personal data in a manner compatible with the research specified in this application.

(d) You will not process this data in ways likely to cause substantial damage or distress to individuals.

(e) You will not use this data to support measures or decisions with respect to individuals (e.g. automated processing or profiling).

(f) You will not share any personal data outside of the research and supervisory team.

(g) Wherever possible, you will pseudonymise data and keep personal data encrypted and separated from research data, so that it is impossible to identify individuals from the research data.

(h) You will ensure research data is fully anonymised, using UK Data Service guidelines, before sharing it.

(i) You will not transfer Personal Data originating from inside European Economic Area (EEA) outside the EEA.

(j) If you need to share personal data with an external organisation who is authorised to access the data, you will use UCL Dropbox service and you will *not use your personal* Dropbox or email account.

(k) You will secure your computer and lock it when not in use to ensure other people cannot see the Personal Data (e.g. by people looking over your shoulder in when in shared spaces).

(l) You will store any electronic Personal Data on your N: Drive personal storage space (with up to 100GB available, for more storage space contact your supervisor and ISD) or on encrypted portable devices (mobile phones, laptops, USB flash drives, or portable hard drives), using UCL Information Security guidance.

(m) You will store any manual Personal Data in locked units – from when participants submit until it is securely destroyed.

(n) You will be responsible for performing regular backups of the data.

(o) You will not keep the Personal Data for any longer than is strictly necessary.

(p) You will securely destroy Personal Data items when it is no longer required.

☐ Please confirm that you have thoroughly read and understood these practices and that you will apply them throughout the course of your research.

6.4. Will you collect or process data on the following *potentially sensitive* topics?

(a) Physical health or condition
(b) Mental health or condition
(c) Use of health care services
(d) Sex life
(e) Sexual orientation
(f) Religious / similar beliefs
(g) Racial or ethnic origin
(h) Political opinions
(i) Trade union membership
(j) Criminal record / commission / alleged commission of an offence, or related proceedings / sentencing
(k) Genetic or biometric data
(l) Other questions participants could find sensitive/upsetting (not covered under Data Protection law)

◉ Confirm that **you are not collecting and processing data** on any of the potentially sensitive topics listed above

– OR –

○ For **each** of these potentially sensitive categories, please justify why you need them and explain how the risk of processing these data are balanced against the public benefit of your research.

Remember that:

- *You may only collect or process this data if it is in the public interest for you to do so.*
- *You may only collect or process this data if you are collecting the minimum sensitive data necessary.*
- *You may only collect or process this data if appropriate safeguards are in place.*

- *You must NOT store electronic sensitive personal data on portable devices such as mobile phones, laptops, USB flash drives, or portable hard drives.*
- *You should ONLY store electronic sensitive personal data on your N: Drive (part of the Filestore@UCL central file storage service).*

7. Declaration and next steps

I confirm that:

- ☐ The information provided is accurate to the best of my knowledge.
- ☐ I will begin with my research only after I have received ethical approval.
- ☐ If answers to any of these questions change, I will submit a new ethics application and secure approval before I begin with my research.
- ☐ If applicable, I have attached local ethics approval from overseas.
- ☐ If applicable, I have attached a risk assessment.
- ☐ If applicable, I have attached the full set of questions / interview guides.
- ☐ If applicable, I have attached the Participant Information and Consent Sheet.
- ☐ If applicable, I have completed the UCL Data Protection training.

8. Supervisor sign-off

Please send this completed form to your supervisor for sign-off.

Supervisors only:

☐ I confirm that all information provided in this form is accurate and that I fully support this application.

Supervisors: please return form to student.

9. Next steps

Please submit your form to Moodle.

The form will be reviewed by the Departmental Ethics Reviewers according to the evaluation scheme shown on the next page. The outcome of the evaluation will be shared on Moodle. You must await the outcome of the evaluation before you can begin your research.

Reviewer checklist

Section 2 – Research methods

- The project is described clearly.
- The methods described seem feasible and adequate for the research project.

Section 3 – Research location, permissions and risks to researchers

- It is clear whether research will be conducted in a country requiring local ethical approval and, if so, sufficient evidence of local ethical approval has been provided.
- It is clear whether permission to conduct the research is necessary and, whether this will be obtained prior to data collection.
- Any risks to the researcher have been clearly identified and, if needed, evidence of risk assessment has been provided.

Section 4 – Details of participants (Refer to attached participant Information and consent form

- The processes of choosing and inviting participants and providing participant information in advance are all clearly stated.
- A Participant Information and Consent Sheet has been developed and covers all relevant points, including contact details, funder, details of the study or experiment, potential risks/harms, anonymity/confidentiality, voluntariness, right to withdraw, data protection, participant declaration.
- The Participant Information and Consent Sheet is written in an appropriate style
- The process of recording consent uses an approved method listed in question 4.2.

Section 5 – Accessing and processing pre-collected data

- The process of obtaining pre-collected data is clear and plausible.
- The data sources and ownership are clearly stated.
- Permission to access them has been obtained. If not, a valid justification is provided.
- If data will be analysed for a different purpose than the original remit, a ethical issues have been identified and will be addressed effectively.

Section 6 – Personal Data and Sensitive Data

- Applicant correctly identifies whether personal data are being collected or processed.
- Applicant correctly identifies whether potentially sensitive data are being collected / processed.
- Any uses of Sensitive Data are sufficiently justified and reflected on with regard to the public benefit of the research.
- Applicant confirms that Personal Data or Sensitive Data will be processed according to the fair processing practices.
- If sensitive questions or data are used, all questions, discussion or interview guides are attached, if applicable.

Appendix B Sound Level Sensor Physical Image and Sensor Specification

Features

- EU, US frequency support
- LoRaWAN 1.0.3
- MEMS microphone
- Resolution: 1 dB
- Measuring range: 40~100dB(A)
- Frequency Response: 100~10,000 Hz
- 0 to 50°C operation temperature range

The Sound Level Sensor utilizes LoRaWAN connectivity to provide to easily measure and investigate sound levels in decibels (dBA) in a variety of building environments.

Specification	
Frequency	868MHz & 915MHz
RF TX power	US: +19dBm EU: +17dBm
Antenna Gain	-2dBi Peak, -5dBi Avg
RF Sensitivity	-135dBm
Battery Type	3.6V ½ AA Li-SOCl2, 1200mAh
Battery Lifetime*	Up to 16 months
Average Current	135mA maximum / 100uA minimum
Protocol	LoRaWAN 1.0.3
Operation Temperature	0 to 50°C
Environmental Rating	IP40 equivalent
Dimensions	50mm x 20mm x 50mm
Weight	30g without battery 40g with battery
Measuring range	40~100dB(A)
Frequency Response	100~10,000 Hz
Type Approval	FCC/IC/CE

Appendix C Activity-Related Noise Sensitivity and Maximum Thresholds for Indoor Ambient Noise Level (Education Funding Agency)

The table is taken from page 17 of BB93: Acoustic design of schools - performance standards, published by the Education Funding Agency.

Type of room	Room classification for the purpose of airborne sound insulation in Tables 3a and 3b		Upper limit for the indoor ambient noise level $L_{Aeq,30mins}$ dB	
	Activity noise (Source room)	Noise tolerance (Receiving room)	New build	Refurbish-ment
Study room (individual study, withdrawal, remedial work, teacher preparation)	Low	Medium	40	45
Libraries:				
Quiet study area	Low	Medium	40	45
Resource area	Average	Medium	40	45
Science laboratory	Average	Medium	40	45
Design and technology:				
Resistant materials, CADCAM area	High	High	40	45
Electronics/control, textiles, food, graphics, design/resource area, ICT room, art	Average	Medium	40	45
Drama studio, assembly hall, multi-purpose hall (drama, PE, audio/visual presentations, assembly, occasional music)	High	Low	35	40
Atrium, circulation space not intended for teaching and learning	Average	Medium	45	50
Sports hall Dance studio Gymnasium/Activity studio	High	Medium	40	45
Swimming pool	High	High	50	55
Meeting room, Interviewing/counselling room, video conference room	Low	Medium	40	45
Dining room	High	High	45	50

Appendix D Guideline Values for Community Noise in Specific Environments Table (WHO)

This document is taken from page 16 of Guideline for Community Noise, the outcome of a meeting publicised by the World Health Organization in 1999.

Specific environment	Critical health effect(s)	L_{Aeq} [dB(A)]	Time base [hours]	L_{Amax} fast [dB]
Outdoor living area	Serious annoyance, daytime and evening Moderate annoyance, daytime and evening	55 50	16 16	- -
Dwelling, indoors Inside bedrooms	Speech intelligibility & moderate annoyance, daytime & evening Sleep disturbance, night-time	35 30	16 8	 45
Outside bedrooms	Sleep disturbance, window open (outdoor values)	45	8	60
School class rooms & pre-schools, indoors	Speech intelligibility, disturbance of information extraction, message communication	35	during class	-
Pre-school bedrooms, indoor	Sleep disturbance	30	sleeping-time	45
School, playground outdoor	Annoyance (external source)	55	during play	-
Hospital, ward rooms, indoors	Sleep disturbance, night-time Sleep disturbance, daytime and evenings	30 30	8 16	40 -
Hospitals, treatment rooms, indoors	Interference with rest and recovery	#1		
Industrial, commercial shopping and traffic areas, indoors and outdoors	Hearing impairment	70	24	110
Ceremonies, festivals and entertainment events	Hearing impairment (patrons:<5 times/year)	100	4	110
Public addresses, indoors and outdoors	Hearing impairment	85	1	110
Music and other sounds through headphones/ earphones	Hearing impairment (free-field value)	85 #4	1	110
Impulse sounds from toys, fireworks and firearms	Hearing impairment (adults) Hearing impairment (children)	- -	- -	140 #2 120 #2
Outdoors in parkland and conservations areas	Disruption of tranquillity	#3		

#1: As low as possible.
#2: Peak sound pressure (not LAF, max) measured 100 mm from the ear.

Appendix E Main Function Codes

The core code is shown below, see GitHub for the full code.

- Define data preprocessing functions for LSTM and GRU model

```
def process_data(data, look_back=1):
    data['date'] = pd.to_datetime(data['date'], format='%d/%m/%Y %H:%M')
    data['year'] = data['date'].dt.year
    data['month'] = data['date'].dt.month
    data['day'] = data['date'].dt.day
    data['hour'] = data['date'].dt.hour
    room_encoded = pd.get_dummies(data['room'], prefix='room')
    data = pd.concat([data, room_encoded], axis=1)
    features = data[['year', 'month', 'day', 'hour', 'Value'] + list(room_encoded.columns)]
    target = data['count_1s']

    look_back_data = []
    for i in range(len(features) - look_back + 1):
        t = features.iloc[i:i+look_back].values
        look_back_data.append(t)

    features = np.array(look_back_data)
    target = target.values[look_back-1:]

    return features, target
```

- Core code for LSTM model building and cross validation

```
look_back = 24
features_qian, target_qian = process_data(data_qian, look_back=look_back)
features_zhong, target_zhong = process_data(data_zhong, look_back=look_back)

features = np.concatenate((features_qian, features_zhong))
target = np.concatenate((target_qian, target_zhong))

n_samples, look_back, n_features = features.shape
features = features.reshape((n_samples, look_back * n_features))

scaler = MinMaxScaler(feature_range=(0, 1))
scaled_features = scaler.fit_transform(features)
scaled_features = scaled_features.reshape((n_samples, look_back, n_features))

k_folds = 5
kf = KFold(n_splits=k_folds, shuffle=False)
fold_indices = kf.split(scaled_features)

# Create an empty DataFrame to store cross-validation results
results_df = pd.DataFrame(columns=['Fold', 'MSE', 'RMSE', 'MAE', 'R2'])

for fold, (train_idx, val_idx) in enumerate(fold_indices):
    X_train_fold, X_val_fold = scaled_features[train_idx], scaled_features[val_idx]
    y_train_fold, y_val_fold = target[train_idx], target[val_idx]

    # Create LSTM model
    model_lstm = Sequential()
    model_lstm.add(LSTM(units=50, return_sequences=True, input_shape=(look_back, n_features)))
    model_lstm.add(Dropout(0.2))
    model_lstm.add(LSTM(units=50))
    model_lstm.add(Dropout(0.2))
    model_lstm.add(Dense(units=1))
    model_lstm.compile(optimizer='adam', loss='mean_squared_error')

    model_lstm.fit(X_train_fold, y_train_fold, epochs=100, batch_size=32)

    predicted_values_lstm = model_lstm.predict(X_val_fold)

    mse_fold = mean_squared_error(y_val_fold, predicted_values_lstm)
    rmse_fold = np.sqrt(mse_fold)
    mae_fold = mean_absolute_error(y_val_fold, predicted_values_lstm)
    r2_fold = r2_score(y_val_fold, predicted_values_lstm)
```

- Core code for GRU model building and cross validation

```
look_back = 24
features_qian, target_qian = process_data(data_qian, look_back=look_back)
features_zhong, target_zhong = process_data(data_zhong, look_back=look_back)

features = np.concatenate((features_qian, features_zhong))
target = np.concatenate((target_qian, target_zhong))

n_samples, look_back, n_features = features.shape
features = features.reshape((n_samples, look_back * n_features))

scaler = MinMaxScaler(feature_range=(0, 1))
scaled_features = scaler.fit_transform(features)
scaled_features = scaled_features.reshape((n_samples, look_back, n_features))

k_folds = 5
kf = KFold(n_splits=k_folds, shuffle=False)
fold_indices = kf.split(scaled_features)

# Create a DataFrame to store the results
results_df = pd.DataFrame(columns=['Fold', 'MSE', 'RMSE', 'MAE', 'R2'])

for fold, (train_idx, val_idx) in enumerate(fold_indices):
    X_train_fold, X_val_fold = scaled_features[train_idx], scaled_features[val_idx]
    y_train_fold, y_val_fold = target[train_idx], target[val_idx]

    model_gru = Sequential()
    model_gru.add(GRU(units=50, return_sequences=True, input_shape=(look_back, n_features)))
    model_gru.add(Dropout(0.2))
    model_gru.add(GRU(units=50))
    model_gru.add(Dropout(0.2))
    model_gru.add(Dense(units=1))
    model_gru.compile(optimizer='adam', loss='mean_squared_error')

    model_gru.fit(X_train_fold, y_train_fold, epochs=100, batch_size=32)

    predicted_values_gru = model_gru.predict(X_val_fold)

    mse_fold = mean_squared_error(y_val_fold, predicted_values_gru)
    rmse_fold = np.sqrt(mse_fold)
    mae_fold = mean_absolute_error(y_val_fold, predicted_values_gru)
    r2_fold = r2_score(y_val_fold, predicted_values_gru)
```

- Define hyperparameter loops to find optimal hyperparameters

```
# Define the hyperparameter combinations to try
look_back_values = [24]
units_values = [50, 100, 150]
dropout_values = [0.2, 0.3, 0.4]

# Lists to store the results
results = []

# Perform hyperparameter tuning
for look_back in look_back_values:
    for units in units_values:
        for dropout in dropout_values:
            mse_scores = []
            kf = KFold(n_splits=5, shuffle=False)
            for train_idx, val_idx in kf.split(scaled_features):
                X_train_fold, X_val_fold = scaled_features[train_idx], scaled_features[val_idx]
                y_train_fold, y_val_fold = target[train_idx], target[val_idx]

                model = Sequential()
                model.add(GRU(units=units, return_sequences=True, input_shape=(look_back, n_features)))
                model.add(Dropout(dropout))
                model.add(GRU(units=units))
                model.add(Dropout(dropout))
                model.add(Dense(units=1))
                model.compile(optimizer='adam', loss='mean_squared_error')

                model.fit(X_train_fold, y_train_fold, epochs=100, batch_size=32, verbose=0)

                predicted_values = model.predict(X_val_fold)
                mse_fold = mean_squared_error(y_val_fold, predicted_values)
                mse_scores.append(mse_fold)

            mean_mse = np.mean(mse_scores)
            std_mse = np.std(mse_scores)
            results.append({'look_back': look_back, 'units': units, 'dropout': dropout, 'mean_mse': mean_mse, 'std_mse': std_mse})
```

Appendix F Research Log

This appendix contains documentation of this research process.

DATE	NOTES
MARCH	Research proposal preparing.
MARCH 31ST	Meeting with Valerio to discuss thesis ideas and Ethical Risk Form A.
APRIL	Research Question preparing.
MAY	Define the research question and start the pre-research for the literature review.
MAY 23RD	Meeting with Valerio to clarify the dataset and started the pre-preparation of the literature review.
JUNE	Frame the methodology.
JUNE 1ST	Meeting with Valerio to discuss noise level and data cleaning methods.
JUNE 14TH	Meeting with Valerio, discussed sensor installation, and initial data obtained, expanded on discussion of observational methods.
JULY	Start writing the literature review and methodology and start running sample data.
JULY 6TH	Meeting with Valerio to discuss specific methods of temporal and spatial analysis.
JULY 13TH	Meeting with Valerio, continue to explore methods.
JULY 20	Check the Literature Review with Valerio.
AUGUST	Improving the entire research programme.
AUGUST 3RD	Meeting with Valerio to discuss missing data issues and the forecasting modelling process.
AUGUST 11TH	Check the Methodology and Result with Valerio.
AUGUST 23RD	Check the full document with Valerio.