MECH0020 Individual Project

AY 2022/23

Student:	**Kyoya Higashino**
Project Title:	**Optimising UCL Study Spaces with the Internet of Things and Machine Learning**
Supervisor:	**Prof. Giles Thomas**

Word Count: 7490

University College London
Torrington Place
LONDON WC1E 7JE

Declaration

I, Kyoya Higashino, confirm that the work presented in this report is my own. Where information has been derived from other sources, I confirm that this has been indicated in the report.

University College London
Torrington Place
LONDON WC1E 7JE

Abstract

With more than 51,000 students (1), finding study space at UCL can be a challenge, resulting in lost study hours and a diminished student experience. In response to this issue, UCL has invested heavily in its Internet of Things (IoT) infrastructure to gather valuable data on occupancy patterns. However, current data analysis applications are limited, only providing real-time occupancies of popular study areas. Therefore, to help students find optimal study spaces, an autonomous data analysis system was conceived, utilising historic data and machine learning to provide appropriate study space recommendations for both current and future use.

Primarily, this project revolves around how raw data from UCL's three most popular areas (Main Library, Student Centre, and Science Library) is transformed into comprehensive information for students. To do so, multiple investigations were pursued, such as relative popularity, machine learning regression techniques, feature analysis, and data set size. These investigations allowed for a fully autonomous Python-coded system to be developed, capable of extrapolating historic occupancy trends and predictive machine learning.

Additionally, infrastructure and mobile app concepts are provided to demonstrate the system's feasibility for practical implementation. The infrastructure concept integrates existing IoT systems and additional cloud-computing services to support data collection, the analysis system, and the mobile app. In the mobile app concept, user-flow diagrams and concept art illustrate how results from the data analysis system can be presented in a student-friendly way.

Using UCL's electronic turnstiles to collect data, the data analysis system was successful in extracting historic trends and predicting occupancies with as little as three weeks of training data. However, due to the Student Centre's high popularity, its predictions are slightly less accurate in comparison to other areas, also requiring more training data to achieve minimum accuracy. Combined with the infrastructure and mobile concepts, this project highlights how occupancy data can be used to improve study space usage and enhance the UCL student experience.

All project code is publicly available at https://github.com/khigashinosg/IndividualProject.git.

University College London
Torrington Place
LONDON WC1E 7JE

Acknowledgements

I am deeply grateful to my project supervisor Prof. Giles Thomas and co-supervisor Dr. Andrea Grech La Rosa for their invaluable mentorship and expert guidance. Their encouragement, feedback, and support helped me maintain my motivation and drive throughout the project, and I could not have completed it without their assistance. I would also like to express my gratitude to UCL Security Systems Manager Mike Dawe and UCL Head of Technical Operations Ketan Parikh for their valuable insights and assistance in obtaining occupancy data, which was instrumental in the success of this project. I would also like to extend my thanks to Prof. Duncan Wilson for providing valuable insights into UCL's current state of people-counting and analysis, as well as the human-related issues associated with space optimisation within large organizations. I am thankful for the contributions of each of these individuals, which made this project a great success.

University College London
Torrington Place
LONDON WC1E 7JE

Table of Contents

University College London
Torrington Place
LONDON WC1E 7JE

University College London
Torrington Place
LONDON WC1E 7JE

University College London
Torrington Place
LONDON WC1E 7JE

Nomenclature and Abbreviations

AD	–	Anderson-Darling
AI	–	Artificial Intelligence
API	–	Application Programming Interface
COBF	–	Curve of Best Fit
GDPR	–	General Data Protection Regulations
ID	–	Identification
IoT	–	Internet of Things
ISD	–	Internet Services Division
JSON	–	JavaScript Object Notation
LIDAR	–	Light Detection and Ranging
LOPC	–	Line of Perfect Correlation
LOPP	–	Line of Perfect Prediction
MAE	–	Mean Absolute Error
ML	–	Main Library
MLR	–	Multiple Linear Regression
MSE	–	Mean Square Error
PIR	–	Passive Infrared
RF	–	Random Forest
RMSE	–	Root Mean Square Error
R^2	–	Coefficient of Determination
SC	–	Student Centre
SL	–	Science Library
SVR	–	Support Vector Regression
SW	–	Shapiro-Wilkes
T1	–	Term 1
T2	–	Term 2
UCL	–	University College London

University College London
Torrington Place
LONDON WC1E 7JE

UI – User Interface

UNSW – University of New South Wales

UX – User Experience

University College London
Torrington Place
LONDON WC1E 7JE

List of Tables

University College London
Torrington Place
LONDON WC1E 7JE

List of Figures

University College London
Torrington Place
LONDON WC1E 7JE

University College London
Torrington Place
LONDON WC1E 7JE

University College London
Torrington Place
LONDON WC1E 7JE

Chapter 1 – INTRODUCTION

1.1 The Problem: Study Space Crowding

Academic institutions are responsible for managing various resources to improve the student experience while maintaining operational efficiency (2–4). Despite the complexity of managing large institutions, many organisations are particularly focused on maximising space utilisation (3, 5, 6).

For UCL, a London university with more than 51,000 students (1), study space is particularly scarce and expensive (7) and renders efficient space utilisation a high priority. In large institutions, "the creation of very small efficiencies could compound to significant financial gains, which could then be invested in academic and research priorities and student services" (8). Without space management, these gains are lost, and inefficiencies could lead to a reduction in graduation rates, student access, and research grants (8).

1.2 The Solution: The Internet of Things and People-Counting

While the scarcity of campus spaces has led to the creation of new spaces such as UCL East, UCL's new campus, UCL has invested heavily in space optimisation devices, including motion-detecting under-desk sensors and overhead cameras. These devices form the majority of UCL's Internet of Things (IoT) network, defined as a network of physical devices with embedded Internet communication technology that allows for remote data sharing and analysis.

As IoT technology develops, detecting occupancy and foot traffic in complex environments is becoming easier (9). This field of data analysis is known as 'people-counting,' becoming increasingly popular in academia (10) for its vast applications in real-time monitoring, data analysis, and automation. While people-counting is a single solution, it can address multiple problems faced by universities, including study space management.

University College London
Torrington Place
LONDON WC1E 7JE

In one study from the University of New South Wales (UNSW), academics were able to save 10% of 'room costs' via an AI-based optimisation algorithm that utilised predicted attendance instead of enrolment numbers when allocating classrooms (6). Furthermore, people-counting can allow for better studying and teaching space recommendations while highlighting areas that need further investment to attract students.

1.3 Project Aim: A Study Space Recommendation Algorithm for UCL Students using the Internet of Things and Machine Learning

The project aim is to develop a digital solution that assists UCL students in finding and planning for on-campus studying by leveraging historical occupancy trends and existing IoT infrastructure. Within the project aim, there are primary and secondary deliverables.

Primary:

1. To create a data analysis system capable of extrapolating occupancy trends and using machine learning to provide study space predictions and recommendations.

Secondary:

1. To develop an IoT infrastructure concept that best supports people counting data acquisition at UCL.
2. To propose a mobile app concept that displays the results of the data analysis system, detailing user flow and system-specific functions such as real-time monitoring and future planning.

While infrastructure and mobile app concepts are explored, this project focuses on analysing historical data and applying advanced regression techniques to provide study space recommendations. Therefore, these concepts only outline a commercially feasible infrastructure and an illustration of how data analysis results can be presented in a student-friendly manner.

University College London
Torrington Place
LONDON WC1E 7JE

Chapter 2 – DOUBLE-DIAMOND FRAMEWORK

In this project, a double-diamond framework was used, comprising research and design phases. As seen in Figure 1, each phase consists of an expansion of initial research or design ideas (i.e., Discover/Develop stages) before consolidating to define design objectives or deliver the final design (i.e., Define/Deliver stages).

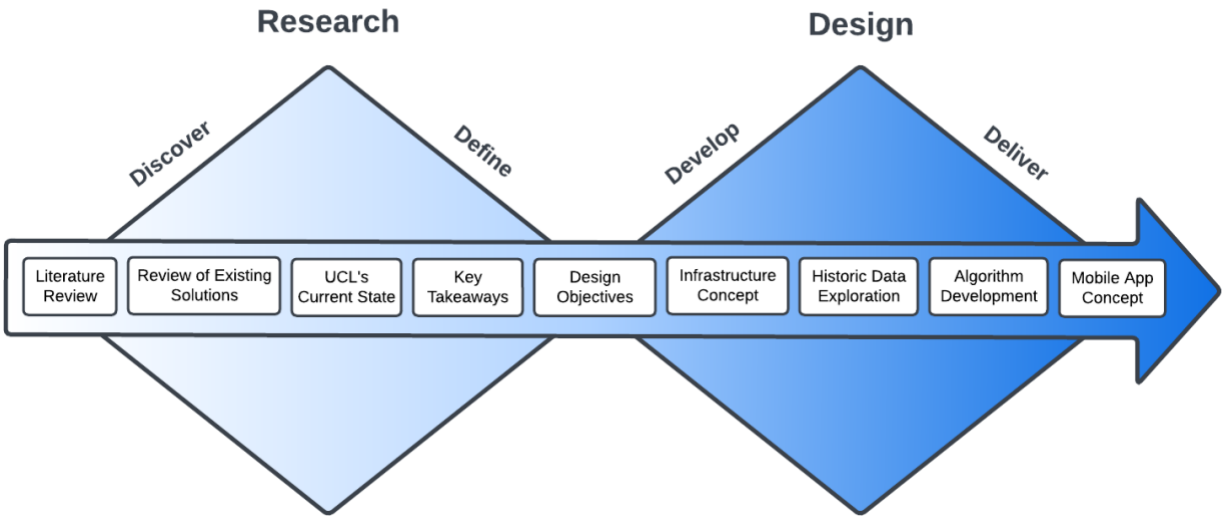

Figure 1: Double-diamond framework comprising project-specific stages and milestones

This double-diamond framework has guided the project workflow, using problem-oriented and UCL-specific research to define design requirements before developing the digital system's three constituent components: infrastructure, data analysis and mobile app concepts. The project-specific activities (white textboxes) are further detailed in this report, with each stage building upon the preceding results to deliver a system design that is detailed and completely relevant.

University College London
Torrington Place
LONDON WC1E 7JE

Chapter 3 – RESEARCH

3.1 Literature Review

3.1.1 Machine Learning Regression Methods

In traditional regression analysis, relationships between dependent and independent variables are established using statistical assumptions and predefined equations (11). With machine learning, AI-based algorithms can automatically learn and make predictions on data, able to handle complex non-linear relationships and improve performance over time (12). A subset of machine learning, supervised learning involves labelled data sets that clearly establish inputs (features) and outputs (responses) to build the regression model. Popular supervised regression methods used in predictive occupancy academia include:

1. Multiple Linear Regression (MLR)
2. Random Forest (RF)
3. Support Vector Regression (SVR)

In MLR, a linear relationship between multiple independent variables and a dependent variable is established, being a better choice for linear modelling (6). Conversely, RF is a non-linear ensemble learning method that aggregates many decision trees to improve predictive accuracy and mitigate overfitting, being better for modelling complex nonlinear relationships (13). Likewise, SVR is also a non-linear method, utilising a hyperplane to map the input space onto a higher-dimensional feature space to find the optimal solution and works well with outliers (14).

3.1.2 Allocating Classrooms using Machine Learning

In a UNSW study, IoT sensors and artificial intelligence (AI) were used to allocate classrooms more efficiently (6). As student lifestyles become more diverse and online content more prevalent, researchers recognised that using predicted attendance instead of traditional enrolment figures is

University College London
Torrington Place
LONDON WC1E 7JE

more suitable for falling attendance numbers. To build their models, three regression methods were tested: MLR, RF, and SVR.

Of the three methods, RF and SVR were the most successful in predicting attendance within the same term, demonstrating the lowest aggregate errors. However, when predicting attendance for the following term, accuracy decreased in all models. Notably, the RF model still performed better in comparison to MLR and SVR.

Despite this decrease in accuracy, researchers were still able to apply the prediction algorithms to save "10% in room costs with a very low risk of room overflows" for 9 different classrooms. While successful, student discomfort due to additional crowding also increased, requiring prediction margins/buffers to mitigate this issue.

3.1.3 Ethical and Societal Impact of Space Optimising Projects

While space optimisation has tremendous benefits, there are significant ethical and societal concerns, specifically regarding crowding and data privacy.

Inextricably linked to space optimisation, crowding is an unavoidable consequence when allocating study space areas. In an interview with Prof Duncan Wilson (Connected Environments, Bartlett Centre for Advanced Spatial Analysis, UCL), he identified that increasing the utilisation of unpopular areas (even without over-crowding) can result in student dissatisfaction among those who used to study there regularly.

Regarding data privacy and consent, Northeastern University in the United States recently received backlash from students and faculty after under-desk occupancy sensors were installed without their consent, resulting in the removal of sensors in protest (15). In a University of Melbourne study where students were polled on their acceptance of hypothetical campus-tracking scenarios, the mean rejection rate was 28.5% even with monetary rewards up to $100 (16). In the same study, the most accepted scenario still had a rejection rate of 10+%.

University College London
Torrington Place
LONDON WC1E 7JE

3.2 UCL's Current State

3.2.1 UCL's People-Counting Infrastructure

UCL has four systems: two primary systems that are solely used for people-counting and two secondary systems whose primary functions are non-occupancy related.

The primary systems utilise technology from FM:Systems (previously OccupEye), a workplace management solutions company, and comprise UCL's 'under-desk sensors' and 'flow-counters.' Under-desk sensors utilise passive infrared (PIR) technology to detect motion at individual desks while flow-counters are positioned above doorways, using infrared cameras to count individual entrances/exits. According to K. Parikh of UCL Information Services Division (ISD) (December 2022), annual maintenance costs are approximately £204k.

Figure 2: An under-desk sensor at Science Library

Figure 3: A flow-counter at Darwin B40

In addition to hardware, UCL also utilises FM:Systems' online platform, capable of space and time utilisation analysis.

University College London
Torrington Place
LONDON WC1E 7JE

Figure 4: Screenshot of FM:Systems online data analysis platform

Conversely, UCL's secondary systems are primarily used for security and attendance, comprising an electronic turnstiles system (Gunnebo SpeedStiles) and a card-based attendance tracking system (RegisterUCL). While data from these systems are internally controlled, minimal work has been done by the UCL administration towards conducting and applying useful occupancy analyses.

Figure 5: Gunnebo SpeedStiles at Science Library

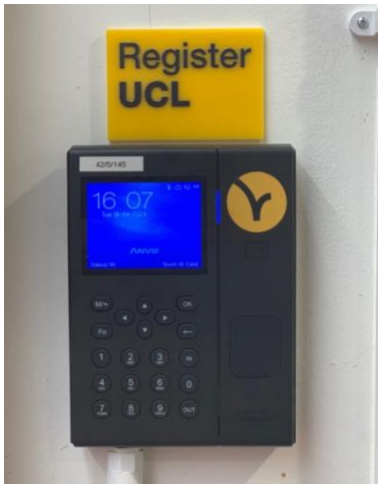

Figure 6: A RegisterUCL card reader at Science Library

Further details about each system are summarised in Table 2.

University College London
Torrington Place
LONDON WC1E 7JE

Table 1: UCL's People-Counting Infrastructure - Details and Advantages

System Name	Technology	Description	Advantages/Disadvantages
FM:Systems Under-Desk Sensors (17)	Passive Infrared Sensors	• Non-identifying, only detecting presence. • Monitors occupancy of singular desks. • Used in most public areas (cluster rooms, study spaces, etc). • Data sent to FM:Systems for analysis and display.	• Non-identifying sensors are General Data Protection Regulations (GDPR) compliant. • Already installed at most popular study spaces. • Platform is highly sophisticated, capable of displaying both space and time utilisation data.
			• Platform only performs historical analysis. • Platform has anti-web scraping protocols and does not allow data downloads.
FM:Systems Flow Counters	Infrared Cameras	• Non-identifying, only detecting presence. • Only entrances and exits are recorded, not infrared camera data. • Covers all 20+ seater classrooms. • Placed above doorways to monitor entrances and exits. • Data sent to FM:Systems for analysis. • Annual maintenance costs are £204k.	• Non-identifying sensors are GDPR compliant.
			• Does not cover study spaces, only lecture halls and classrooms. • Incapable of heat mapping, only overall occupancy. • Difficult to install in many study spaces due to high entrance ceilings.
Gunnebo FP/FPW SpeedStiles (18)	Electronic Turnstiles	• Installed at building entrances/exits. • Requires UCL ID cards for use. • All turnstiles require ID cards for entry. • Only some turnstiles require ID for exit. • Primarily used for security reasons. • Data is controlled by UCL Security, can be exported to Excel.	• Data is internally controlled and easily accessible for advanced occupancy analysis. • Already installed at most popular study spaces.
			• Only overall building occupancy can be obtained, not for specific levels or rooms. • Only individual entrance/exit data is recorded, processing is required to obtain occupancy numbers. • Overrides by security staff result in uncounted entrances/exits, leading to short and long-term errors.
RegisterUCL (19)	ID Card Readers	• Requires UCL ID cards for use. • Covers classrooms and lecture halls. • Primarily used for attendance. • Controlled by UCL Security but is available to all departments to track attendance.	• Not a passive system, requiring card readers. • Does not cover study spaces, only lecture halls and classrooms. • System records personal data, requiring identifiable data to be masked.

University College London
Torrington Place
LONDON WC1E 7JE

3.2.2 UCL's Obstacles

After interviewing Prof. Wilson, it was concluded that UCL's occupancy analysis work is underdeveloped. Overall, four issues need to be overcome before occupancy data can be utilised for study space optimisation at UCL, described below in Table 1.

Table 2: UCL's Four Main Obstacles to Applicable People-Counting

Main Issue	Specific Obstacles
Availability	• Data is not easily available, even to UCL academics. Special permission is needed from UCL to access the FM: Systems platform while SpeedStiles and RegisterUCL data require assistance from UCL Security. • While Application Programming Interfaces (APIs) are currently being developed to help retrieve data, these are still in early development. • Obtaining access to occupancy data is challenging due to conflicting claims from multiple parties, primarily UCL Estates and ISD.
Accuracy	• UCL's primary people-counting systems often break and require maintenance, resulting in a constantly changing total number of sensors and cameras are that being deployed. As occupancy cannot be detected in these affected areas, spatial analysis capabilities are reduced. • Data from UCL's secondary people-counting systems are hardly utilised, having many random and systematic errors.
Occupant Satisfaction	• Space optimisation can result in increased occupancy during previously underutilised times/areas, potentially causing dissatisfaction among students who valued the open nature of these spaces.
Data Privacy and Ethics	• UCL's secondary people-counting systems intrinsically deal with identifiable data, where student profiles are linked to entries and exits. Some students may be against this form of tracking. • While UCL's primary people-counting systems are GDPR compliant, inevitably some students will still be against anonymous tracking.

University College London
Torrington Place
LONDON WC1E 7JE

3.3 Other Existing Solutions

To solve study space crowding on campuses, multiple physical and digital solutions need to be integrated, including other people-counting sensors and mobile apps.

3.3.1 People-Counting Cameras – Terabee

An alternative to FM:Systems, smart sensor company Terabee also supplies people-counting cameras. Instead of infrared technology, Terabee uses time-of-flight light-detection-and-ranging (LIDAR) to count entrances/exits. During self-conducted testing, these sensors proved to be easy to self-install and set up for wireless data transmission. However, they require light drilling for installation and a wired power source.

Figure 7: Terabee's 'People-Counting-M' sensor

Terabee offers an 'M' and 'XL' version of its sensor, with the 'XL' version capable of monitoring wider door entrances of up to 15m using multiple devices (20). In an interview with Terabee (November 2022), the unit cost of both models is €250 and €500 respectively with no subscription cost. While Terabee offers an online data analysis platform like FM:Systems, its sensors can send data to any online server using JSON messages, allowing for more customisable IoT applications.

University College London
Torrington Place
LONDON WC1E 7JE

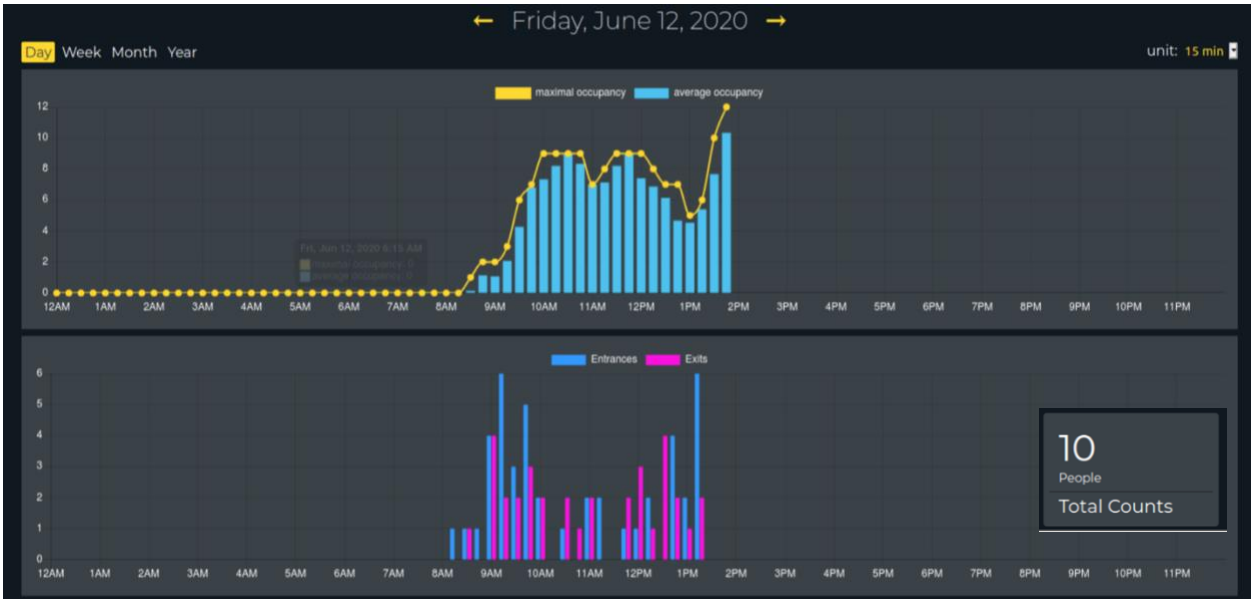

Figure 8: Screenshot of Terabee's online data analysis platform (written communication, November 2022)

3.3.2 Mobile Apps – UCL Go! and Waitz

After occupancy data is processed, recommendations need to be displayed quickly and conveniently. UCL's student app 'UCL Go!' can display real-time occupancies using under-desk sensors but is unable to provide historical trends or future recommendations. Alternatively, several universities have incorporated Wi-Fi tracking technology from IoT solutions provider Occuspace, displaying data through their Waitz app. While Waitz can display both historic and real-time occupancy, it cannot provide future recommendations.

University College London
Torrington Place
LONDON WC1E 7JE

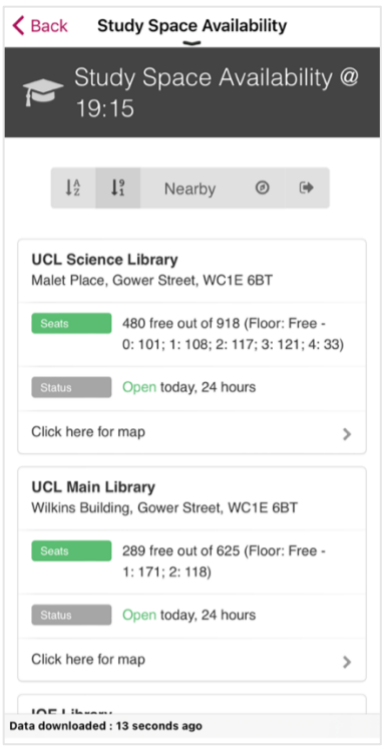

Figure 9: Screenshot of UCL Go! showing real-time occupancy

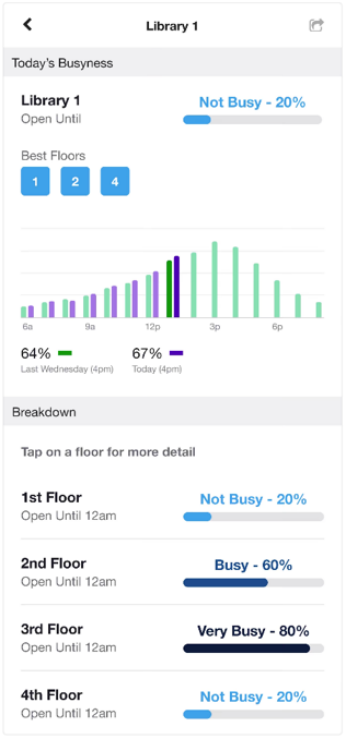

Figure 10: Screenshot of Waitz app, showing historic and real-time occupancy (21)

3.4 Summary and Key Takeaways

Using machine learning to predict occupancy rates has been successful at other universities, having proven applications like reducing costs. However, as machine learning is complex, a thorough investigation is needed to select relevant regression methods and features.

Although UCL's occupancy data analysis is lacking, its developed IoT infrastructure should be utilised. While its primary people-counting systems are more sophisticated, the data from secondary systems are easier to attain and manipulate.

Although most study space apps can display real-time and historic occupancies, none are capable of future predictions or recommendations. Successful apps have a clean user interface (UI) and utilise relevant filters to enhance user experience (UX).

These takeaways allow for the definition of meaningful and attainable design objectives along with measurable success criteria specific to each deliverable.

University College London
Torrington Place
LONDON WC1E 7JE

Chapter 4 – DESIGN

Each deliverable started with the development of clear design objectives and success criteria. Once established, a systematic design process was followed.

University College London
Torrington Place
LONDON WC1E 7JE

4.1 Design Objectives

Table 3: Design Objectives and Success Criteria

Category	Objectives	Success Criteria	Justification
Infrastructure Concept	Utilised data is non-tracking.	GDPR compliant	Solution acceptance increases with a sense of privacy and anonymity.
	Estimated costs are minimal.	Less than FM:Systems annual maintenance costs of £204k.	Should cost less than UCL's current occupancy analysis expenses.
	Incorporates existing UCL infrastructure.	At least 1 system.	Reduces initial investment costs.
Data Analysis System	Full autonomy from raw data to prediction results.	Aside from large random errors, processing raw data requires no manual manipulation.	Solution acceptance increases ease of use.
		Only area, time, and date inputs required for predictive capabilities.	
	Predictions are accurate and consistent.	'Accurate' is defined as within ±15% of the true occupancy for individual predictions.	Occupancy predictions can tolerate some inaccuracy, thus these thresholds balance variability with reasonable accuracy/consistency.
		'Consistent' is defined as having 'accurate' predictions 85% of the time.	
Mobile App Concept	Displays historic, real-time, and predictive occupancy data	Displays real-time data available through UCL Go!	Ensures parity with existing app solutions.
		Displays historic and predictive results from the data analysis system	Improves upon existing app solutions.
	Concept is user-friendly and intuitive.	Text is minimal, large, and clear.	Quick and easy to read.
		Contrasting colour scheme.	
		Inputs are used to display only relevant data (area, time, date).	Allows for custom study space recommendations.

University College London
Torrington Place
LONDON WC1E 7JE

4.2 Design Process

First, an infrastructure concept is established to define the nature of occupancy data and allow data extrapolation to begin, processing raw data to generate historic trends. Next, the prediction algorithm is developed, training regression models on the historic data while conducting investigations such as feature analysis and relative occupancy. Finally, a mobile app concept is designed to display occupancy results while considering UI/UX features.

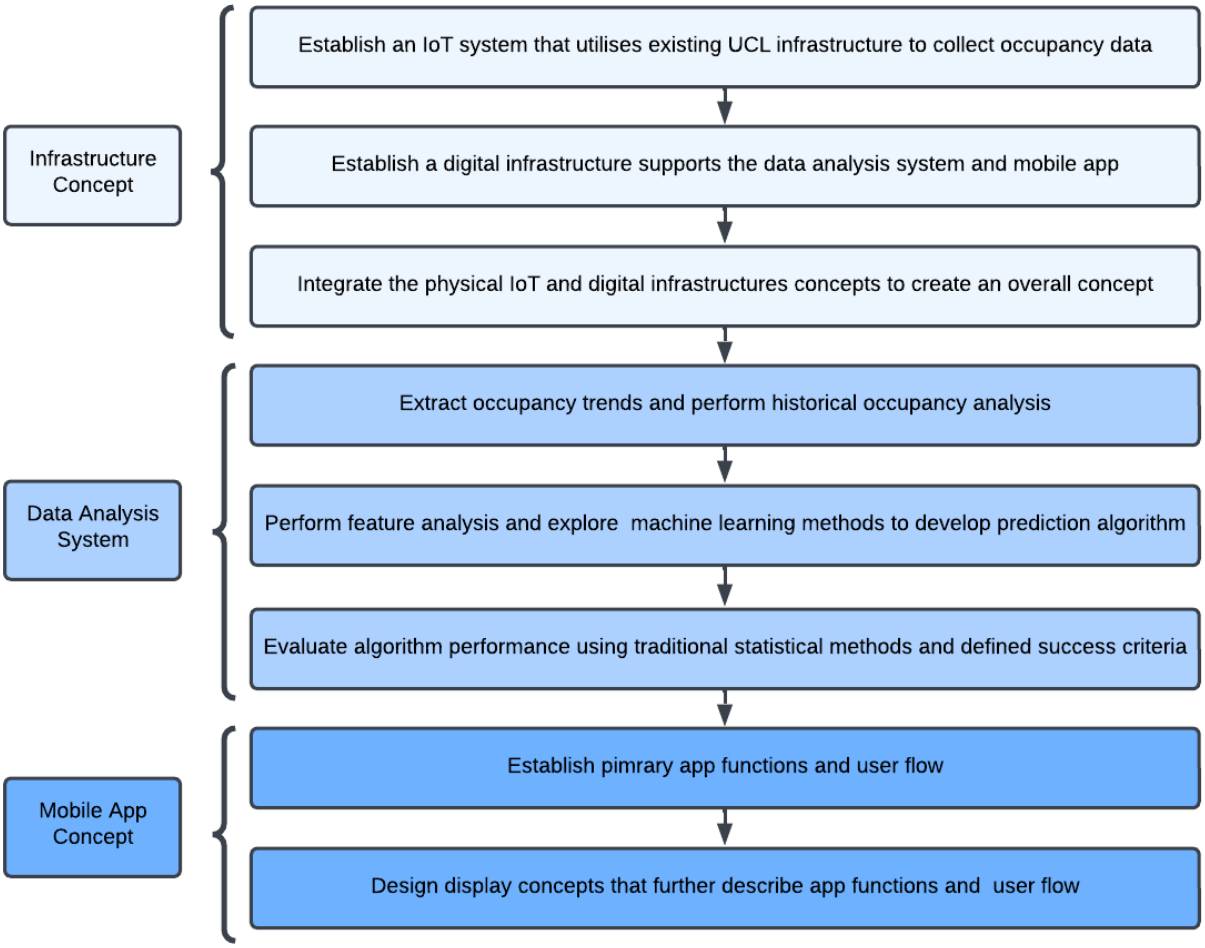

Figure 11: Overall design process flowchart

University College London
Torrington Place
LONDON WC1E 7JE

Chapter 5 – INFRASTRUCTURE CONCEPT

To support data analysis, the infrastructure concept must incorporate physical and digital systems, having data collection and computing capabilities.

5.1 Physical Infrastructure

After exploring multiple sensor systems, the chosen IoT infrastructure is UCL's Gallagher SpeedStiles, being an existing and internally controlled system. Although UCL has other people-counting systems, these are inadequate as the flow-counters and RegisterUCL only monitor classrooms and the under-desk sensors have data extraction difficulties. While the SpeedStiles do record identifiable data, this data is already gathered by UCL Security. When fed into the data analysis system, student ID numbers are masked, thus upholding the privacy of individuals.

5.2 Digital Infrastructure

To support data analysis and the mobile app, a cloud computing service is proposed, such as Google Firebase or Amazon Web Services. Cloud services are modular and flexible, allowing for "pay-as-you-go" subscription packages and quick system upgrades while keeping costs low (22, 23). The combination of both the physical and digital infrastructures is displayed in Figure 12.

University College London
Torrington Place
LONDON WC1E 7JE

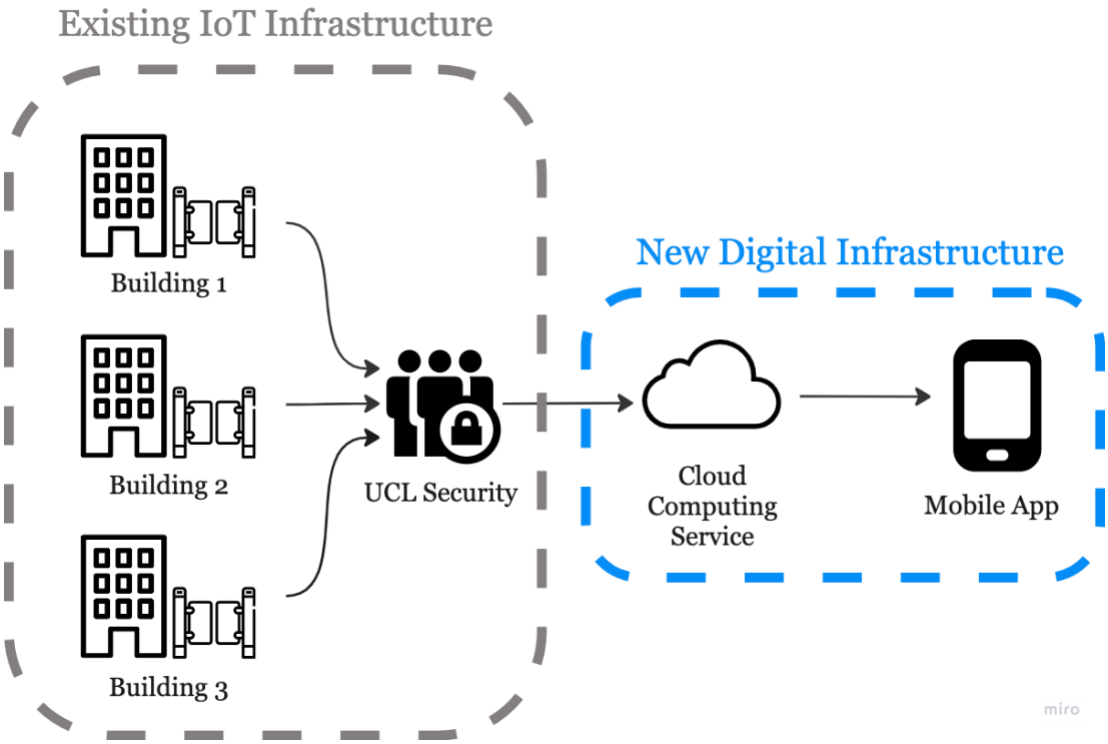

Figure 12: Infrastructure concept utilising existing IoT systems and new cloud computing services

5.3 Cost Analysis

Estimating maintenance costs is difficult without any app development. However, costs should be low considering that the intended user audience is small (UCL students) and occupancy data is not memory intensive, resulting in reduced server, memory, and disk space requirements (24). In 2023, 'ballpark' maintenance costs are £200 – £400, resulting in approximately £5000 in estimated annual costs (25).

University College London
Torrington Place
LONDON WC1E 7JE

Chapter 6 – DATA ANALYSIS SYSTEM

Given the time-consuming nature of data requests, the project focused on UCL's three largest study areas. Further information can be found in Table 4 (26).

Table 4: Analysed Study Areas and Relative Information

ID	Area Name	Maximum Occupancy	Opening Hours	
			Weekday	Weekend
ML	Main Library	625	8:30 – 00:00	11:00 – 21:00
SC	Student Centre	641	24 hours	24 Hours
SL	Science Library	918	Mon: 8:45 – 00:00 Others: 24 hours	Sat: 00:00 – 21:00 Sun: 11:00 – 21:00

Data were collected over the span of Term 1 and most of Term 2, detailed in Table 5. For the full 2022/2023 academic calendar, see Appendix A.

Table 5: Duration of Data Collection

ID	Name	Academic Calendar Duration	Data Collection Duration	Reading Week Dates
T1	Term 1	26 Sep – 16 Dec (2022)	26 Sep – 16 Dec (2022)	7 Nov – 11 Nov (2022)
T2	Term 2	9 Jan – 24 March (2023)	9 Jan – 05 March (2023)	13 Feb – 17 Feb (2023)

6.1 Historic Data Extrapolation

As only entries/exits are recorded, Python code was developed to extract historic occupancy. However, SpeedStiles data is intrinsically prone to random and systematic errors due to breakdowns, scheduled maintenance, and overrides by UCL security. This can be seen in Figure 13, using T1 ML data as a continued example.

University College London
Torrington Place
LONDON WC1E 7JE

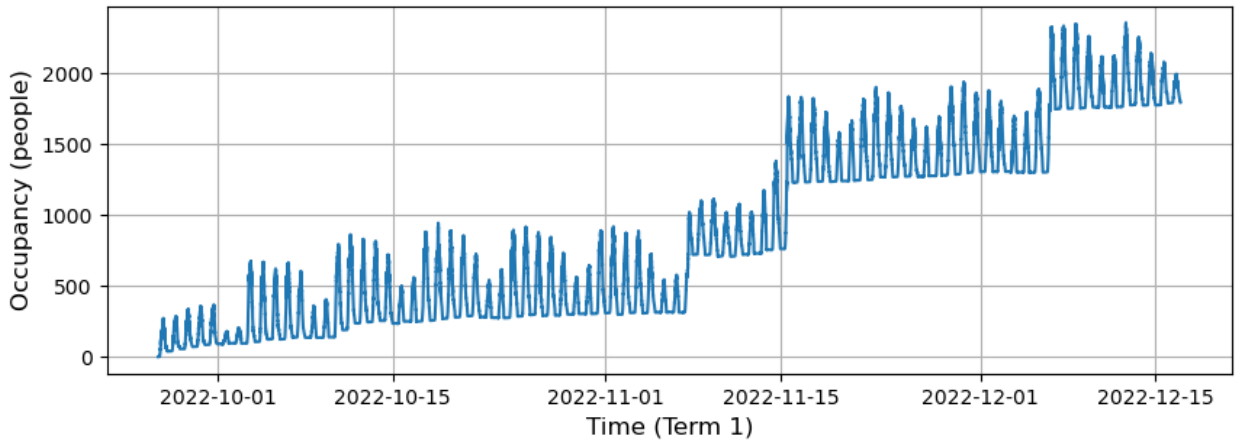

Figure 13: Uncorrected occupancy (T1 ML)

6.1.1 Random Error Correction

These random errors are unusually high net influxes of 200-500 people and are corrected by selectively removing data points according to the error severity. Considering the infrequency and subjective definition of random errors, correction procedures were carried out manually. As in Figure 13, the first random error occurred on Nov 11th, requiring the manual correction seen in Figure 14.

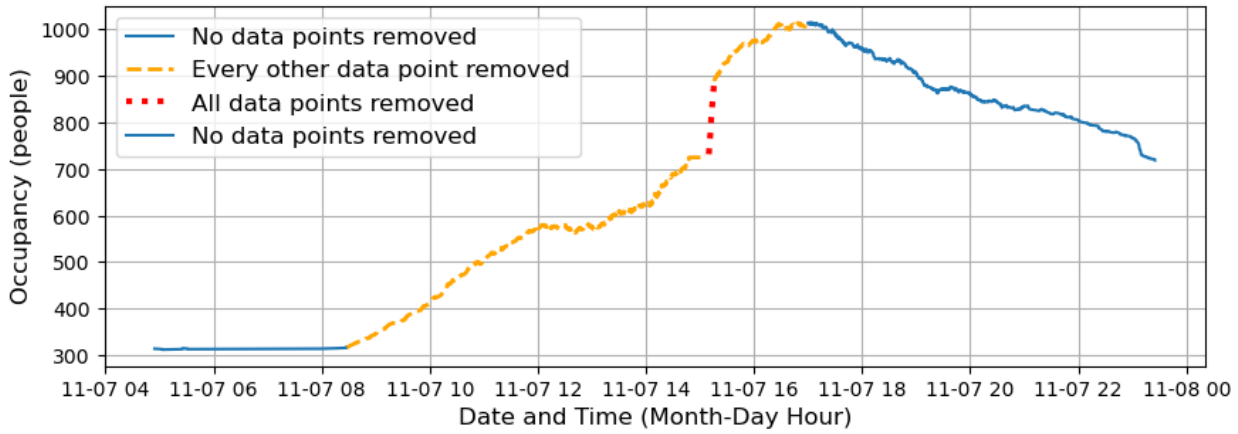

Figure 14: Random error correction example (Nov 7th, ML T1)

University College London
Torrington Place
LONDON WC1E 7JE

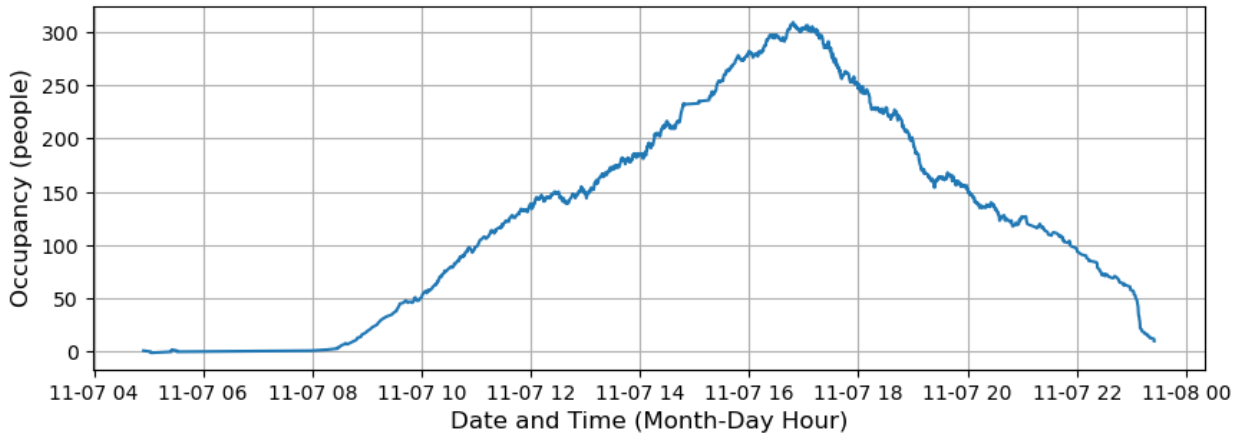

Figure 15: Corrected random error example (Nov 7th ML T1)

6.1.2 Systematic Error Correction

Once random errors are corrected, only systematic error remains, as seen in Figure 16.

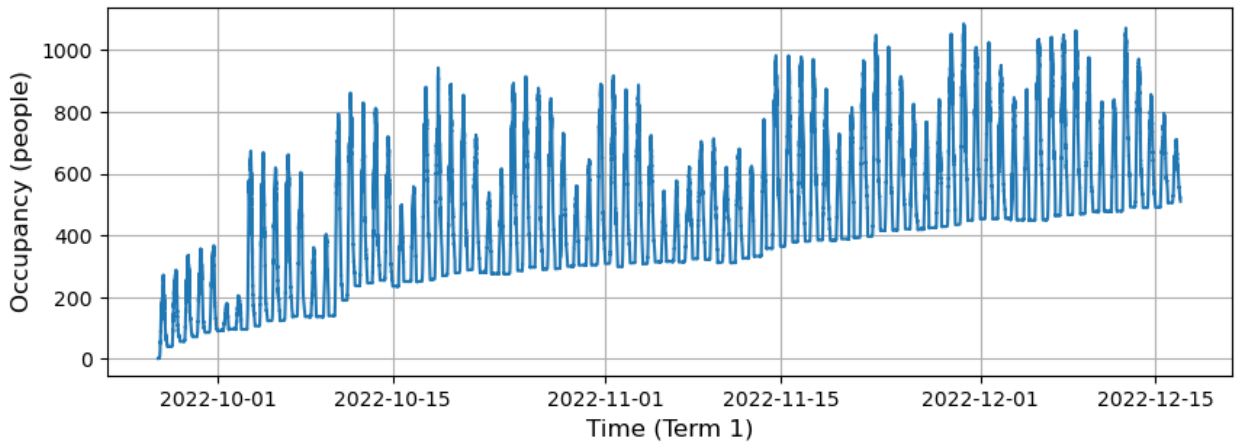

Figure 16: Long-term systematic error example (ML T1)

After testing with linear and polynomial offsets, a daily reset method was chosen, resetting occupancy to 0 at closing time. If an area is always available, occupancy is reset to the average minimum occupancy at the corresponding time as given by UCL's under-desk sensors. While the

University College London
Torrington Place
LONDON WC1E 7JE

under-desk data does not perfectly match the SpeedStiles data, it provides a sufficiently close estimation, also matching anecdotal evidence from students. Hence, the SC and SL are reset to 27 and 10 at 05:00 and 04:00 respectively.

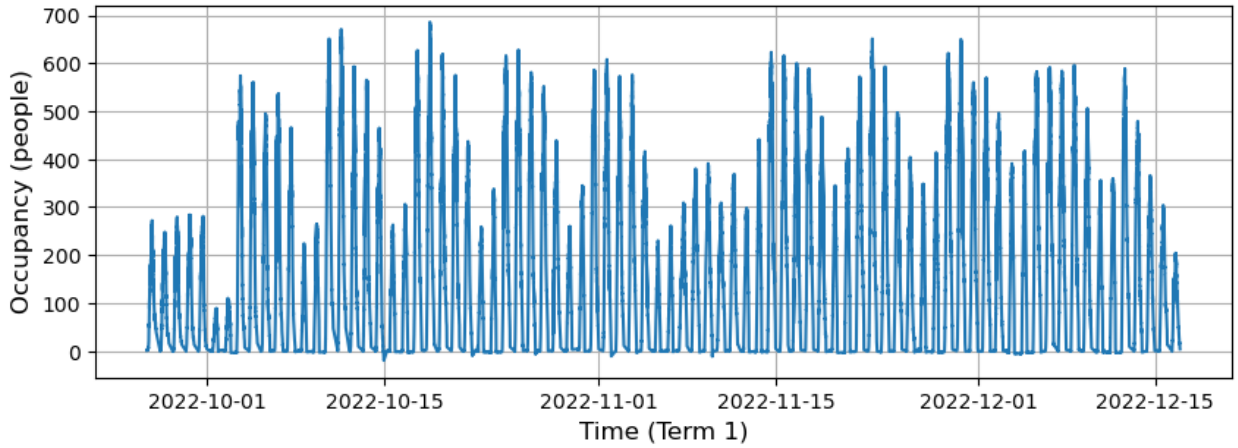

Figure 17: Fully corrected occupancy data example (ML T1)

6.1.3 Daily Trends

Once fully corrected, average daily occupancy trends were calculated (Figure 18).

University College London
Torrington Place
LONDON WC1E 7JE

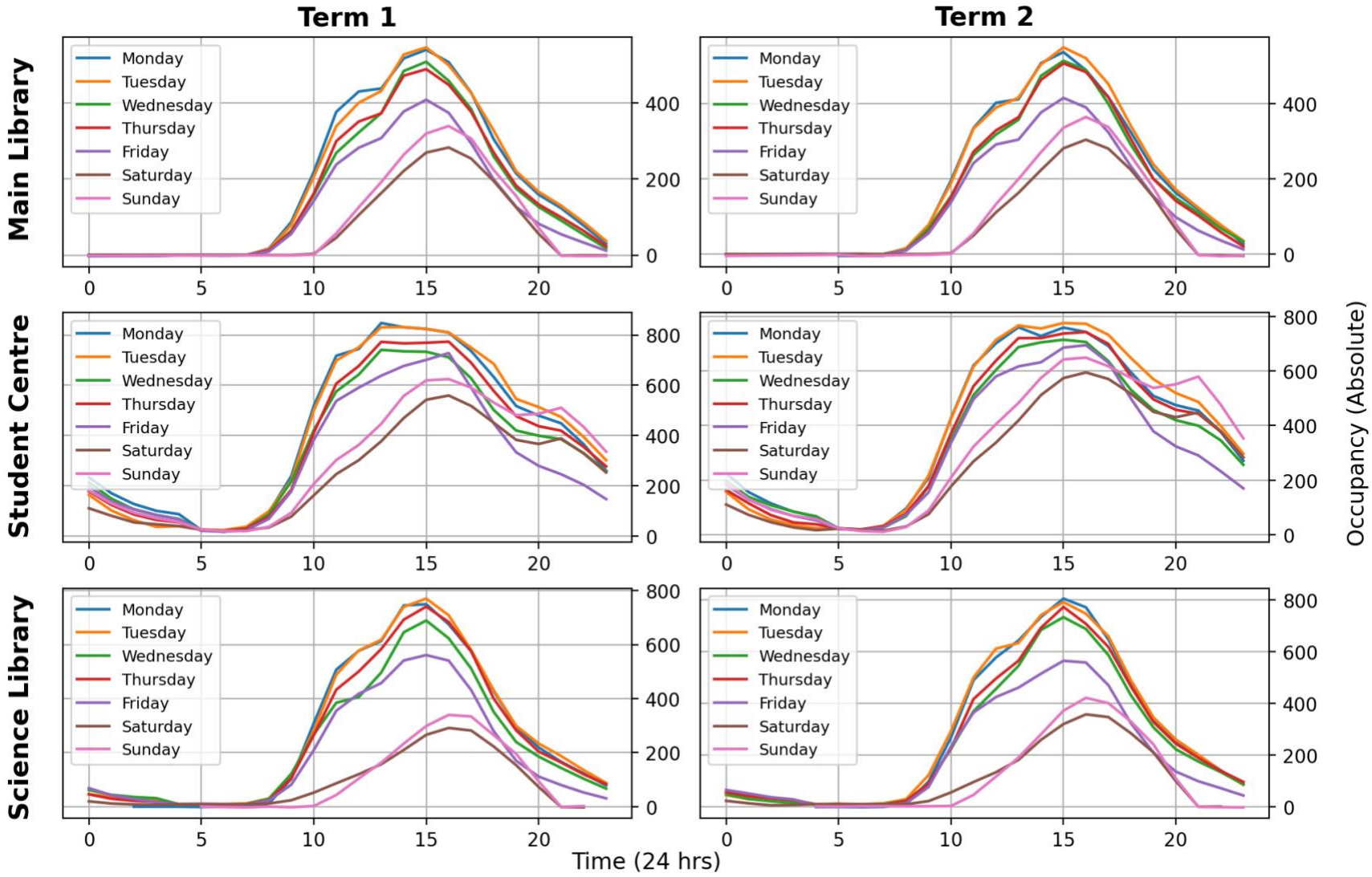

Figure 18: Daily trends of ML, SC, and SL (T1 and T2)

University College London
Torrington Place
LONDON WC1E 7JE

Figure 18 shows that occupancy is dependent on both the time and day of the week. For all areas, Monday and Tuesday are the busiest while Wednesday and Friday are less busy due to a half-day schedule and proximity to the weekend. Expectedly, weekends are the least busy overall.

Differing from ML and SL, the SC's trends are the most unique, having longer peaks and shorter troughs while having minimal decreases in popularity over the weekend. These differences illustrate the importance of the time and day as features in the prediction algorithm for each area.

6.2 Data Exploration and Testing

6.2.1 Relative Occupancy

As data is recorded upon an entrance/exit, data is recorded at irregular intervals. Hence, the average occupancy was taken in 10-minute intervals to regularise and condense datasets. By choosing 10-minute time steps, a balance between accuracy and computational speed was achieved.

When recommending study spaces, the relative crowding of each area rather is more important than absolute occupancy. While large areas may have more seats available, a better choice would be any less densely packed area. Hence, comparative scatterplots were created to understand the occupancy distribution between study areas during shared opening hours. These scatterplots contain a ‘line of perfect correlation’ (LOPC) and a ‘curve of best fit’ (COBF) to illustrate the difference between perfect assumptions and real trends. By comparing the LOPC and COBF, several inferences about the relative popularity of each area can be made.

In Figure 19, the COBF is curved toward the SC axis, showing that students strongly prefer the SC at low to moderate occupancy levels (approx. 10%-60% occupancy). However, at higher occupancy levels (70%+), the COBF returns to the LOPC, signalling that as the SC become more crowded, students are more likely to resort to alternative spaces. When comparing the ML and SL, COBFs closely follow the LOPC in both T1 and T2, displaying very weak preferences for the ML at low occupancies and for the SL at high occupancies. Therefore, a conclusion on the relative popularity of each area can be formed.

University College London
Torrington Place
LONDON WC1E 7JE

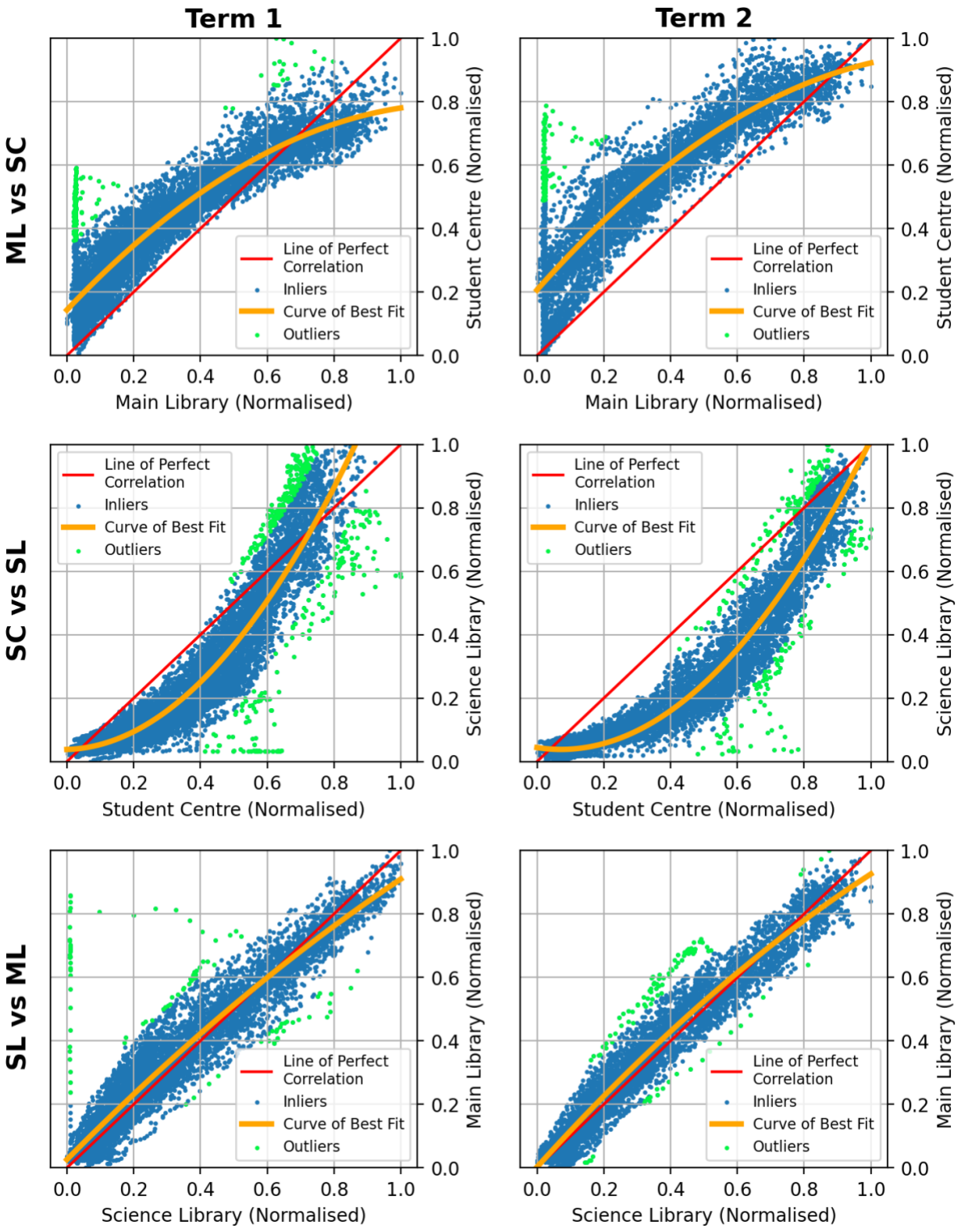

Figure 19: Comparative scatterplots of ML, SC, & SL occupancies (normalised)

University College London
Torrington Place
LONDON WC1E 7JE

Table 6: Relative Popularity of Study Areas

Relative Popularity	Area	Preference Strength
1	Student Centre	Very strongly preferred
T-2	Main Library	Least preferred (tied)
T-2	Science Library	Least preferred (tied)

Outliers have also been highlighted in green, using 2.5 standard deviations from the COBF as the self-imposed threshold. Notably, some outliers are grouped immediately adjacent to a specific axis as seen best in ML vs SC scatterplots. Most of these outliers are adjacent to the 'SC' axis, reflecting moderate SC occupancy when the other areas (ML/SL) are empty. This is due to SC's more reliable opening hours, consistently adhering to its stated opening hours while other areas are known to have unexpected closures throughout the year. Other notable outliers include an unusually dense cluster in Figure 19's SC vs SL (T1) scatterplot, caused by Reading Week. While normally less popular than the SC, these outliers show that students preferred the SL more during this time. This could be due to the SL's high popularity with science-based students (71% of all SL students are from a science-based course) and that science-based students often have mandatory laboratory sessions during Reading Week. Also, since T1 anecdotally presents less academic pressure than T2, non-science students are more likely to avoid campus during this time.

6.2.2 Gaussian Tests

As some machine learning techniques assume normally distributed data, popular Gaussian tests, including the Anderson-Darling (AD) test and histograms, were performed.

The AD test was chosen over others such as the more popular Shapiro-Wilkes (SW) test for its high performance regarding exceptionally large datasets. While one study showed that the SW test slightly outperformed the AD test for datasets up to 2000 points, the best large-data SW algorithm, 'AS R94', should only be used for datasets up to 5000 points (27). As the occupancy dataset sizes range from 5191 to 11773, the AD test was chosen instead.

Through a null hypothesis that assumes a normal distribution, the AD test provides a test statistic as well as critical values that the test statistic must surpass to reject the hypothesis. If rejected, the

University College London
Torrington Place
LONDON WC1E 7JE

data can be categorised as non-Gaussian with a confidence level unique to the critical value used. Below in Table 7 are the AD test statistic results for all six datasets.

Table 7: AD Test Statistics for All Data Sets

	T1	T2
ML	175.74	86.8
SC	339.6	235.3
SL	649.4	340.8

As the 1% significance level is only 1.09, it stated with 99% confidence that all areas follow non-Gaussian occupancy trends. This can also be observed in histogram form, utilising 20 bins to categorise data points.

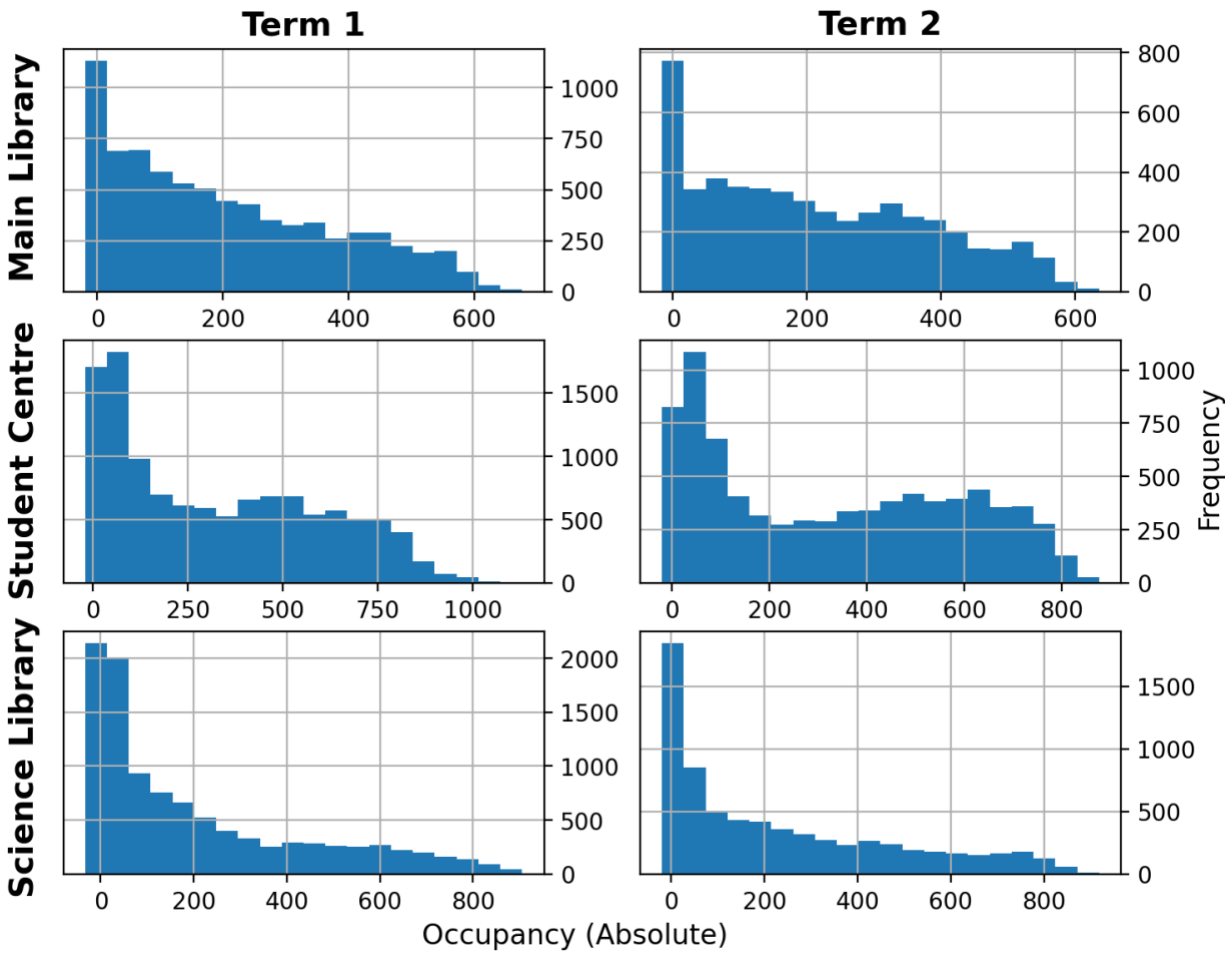

Figure 20: Histograms of occupancy levels during opening hours

University College London
Torrington Place
LONDON WC1E 7JE

6.3 Machine Learning Regression

6.3.1 Pre-processing and Feature Exploration

To build the regression models, the features in Table 8 were explored. All features had to be pre-processed, being continuous or categorical with categorical features being further classified as nominal or ordinal. Nominal features have no intrinsic order while ordinal features do, such as Reading Week (binary) and the Week of Term (ordered).

Table 8: Explored Features for Regression Models

Feature	Description	Type	Numeric Range
Time of Day	Given in 10-minute intervals starting from 00:00.	Ordinal	0 – 143
Day of Week	Monday – Sunday.	Ordinal	0 – 6
Week of Term	With Induction Week, T1 has 12 weeks while T2 has 11 weeks. Induction week is denoted as Week 0.	Ordinal	T1: 0 – 11 T2: 1 – 11
Reading Week	A binary feature. 1 denotes data that falls on Reading Week.	Nominal	0 – 1
Induction Week	A binary feature. 1 denotes data that falls on Induction Week.	Nominal	0 – 1
Occupancy of Alternative Space 1	The occupancy of the first alternative study space.	Continuous	0 – ∞
Occupancy of Alternative Space 2	The occupancy of the second alternative study space.	Continuous	0 – ∞

To determine the correlation strength between specific features and occupancy, Spearman's rank correlation was used, chosen for its non-parametric nature and easily interpretable results. Here, a correlation coefficient ranging from -1 to 1 is produced, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation. For all features, associated p-values are close to 0 (Appendix D).

University College London
Torrington Place
LONDON WC1E 7JE

Table 9: Spearman's Rank Correlation Results (Correlation Coefficients)

Feature	ML		SC		SL	
	T1	T2	T1	T2	T1	T2
Time of Day	0.498	0.507	0.596	0.638	0.493	0.524
Day of Week	-0.217	-0.221	-0.151	-0.122	-0.223	-0.237
Week of Term	0.014	-0.035	0.066	0.021	0.049	-0.017
Reading Week	-0.024	-0.049	-0.039	-0.050	-0.034	-0.062
Induction Week	-0.024	N/A	-0.110	N/A	-0.102	N/A
Alternative Occupancy 1	SC: 0.862	SC: 0.852	ML: 0.861	ML: 0.847	ML: 0.908	ML: 0.936
Alternative Occupancy 2	SL: 0.908	SL: 0.937	SL: 0.858	SL: 0.867	SC: 0.858	SC: 0.872

As seen in Table 9, the occupancies of alternative spaces are the strongest factors in predicting occupancy. While statistically significant, these features cannot be used in practice as it requires knowledge of the future occupancies of alternative spaces. Alternatively, these features could be used for real-time occupancy calculations if an area is unable to determine its current occupancy (e.g., IoT sensors are under maintenance). This investigation is further detailed in Appendix F.

Therefore, the strongest useful feature is the 'Time of Day,' followed by the 'Day of Week' with the remaining features having very weak correlations (< 0.1). However, these features are still included in subsequent machine learning models due to their potential benefit and observable impact on historic trends as seen in Figure 17 and Appendix C.

6.3.2 Method Selection

As introduced in 3.1.1, three machine learning methods were explored: ML, RF, and SVR. In each technique, the chosen regression loss function was the mean square error (MSE) as harsher penalties are desired for large errors (see Appendix E). To evaluate performance, the root mean squared error (RSME) was used to measure the aggregate error while the coefficient of determination (R^2 value) was used to measure the variance in model response. Like the MSE loss function, RSME was chosen to heavily penalise larger errors while R^2 is chosen for its interpretability benefits, ranging from $0 - 1$ instead of $0 - \infty$ like other metrics (28).

University College London
Torrington Place
LONDON WC1E 7JE

To evaluate each method's ability to predict occupancy within the same term and the next, different train-test splits were defined. For same-term occupancy, the training set included T1 data from 26th September to 11th November and was tested on the remainder of T1, resulting in a 57/43 train-test split (Figure 21). For next-term occupancy, models are trained on all T1 data and are tested on T2 data (Figure 22). Like an LOPC, resultant scatterplots also include a 'line-of-perfect-prediction' (LOPP).

As seen in Figure 21 and Figure 22, RF greatly outperforms MLR and SVR with RF's scatterplots closely residing along the LOPP while having the lowest RSME and highest R^2 values.

University College London
Torrington Place
LONDON WC1E 7JE

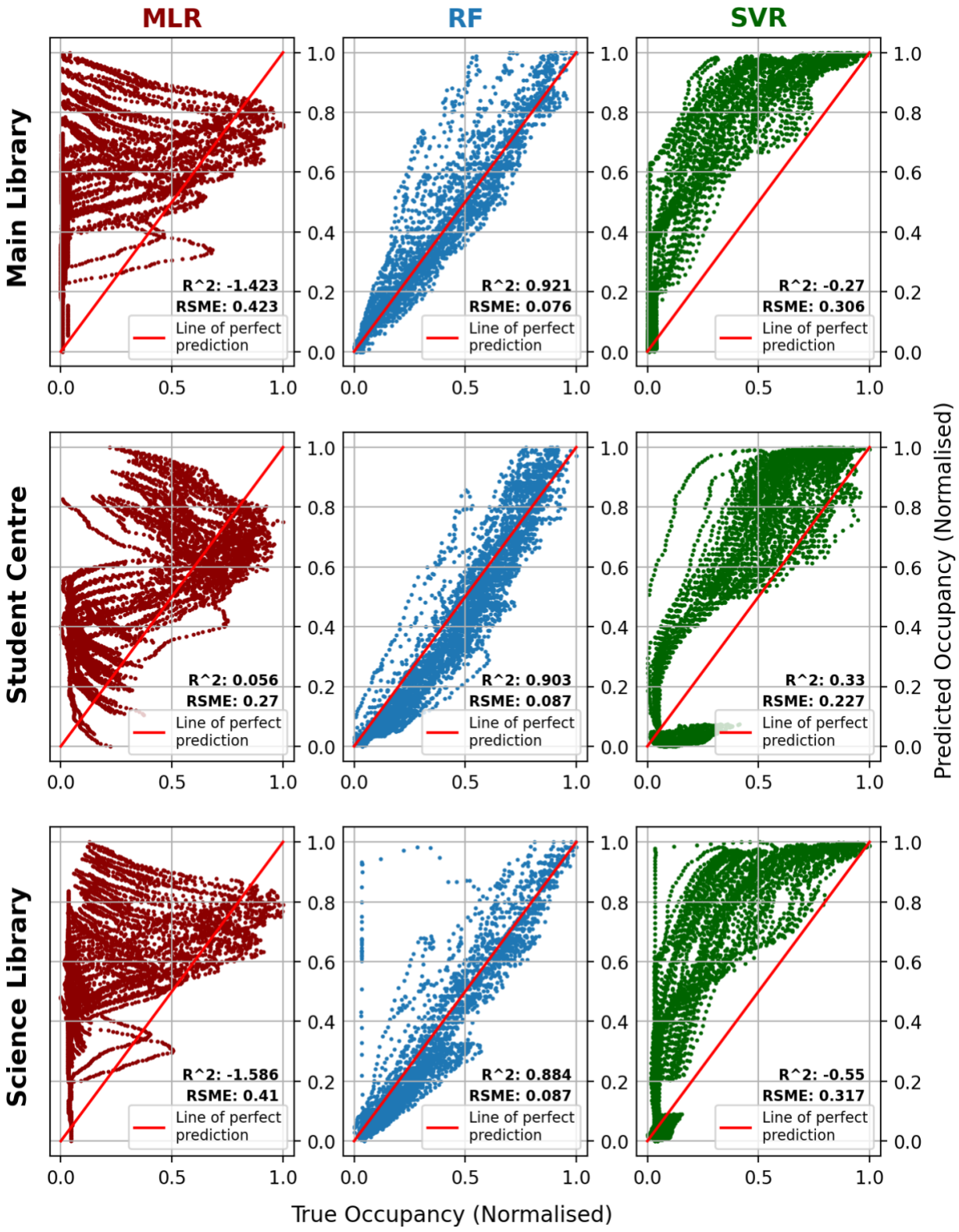

Figure 21: T1 predicted vs. true occupancy (Weeks 1-7 train, 8-12 test)

University College London
Torrington Place
LONDON WC1E 7JE

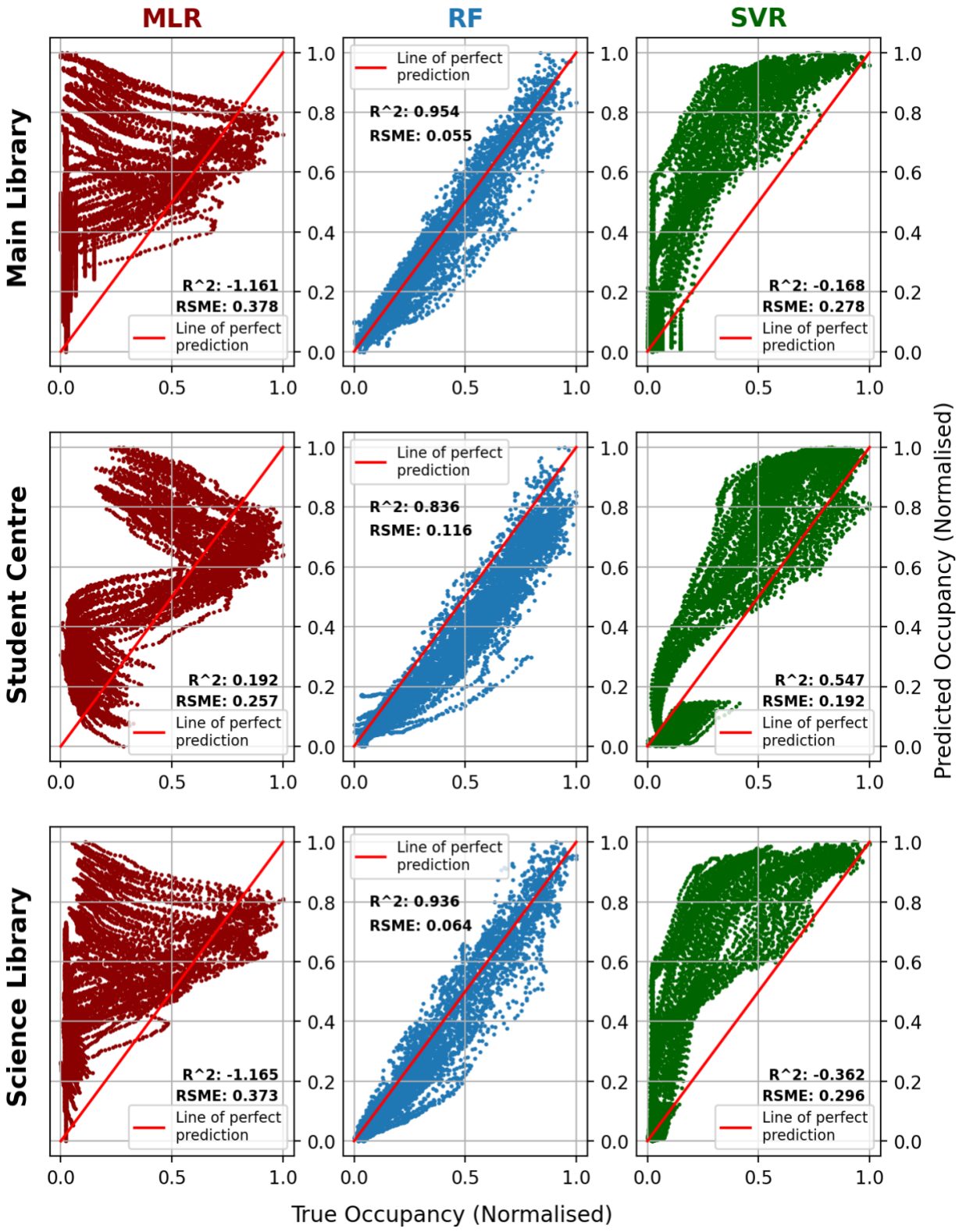

Figure 22: T2 predicted vs. true occupancy (T1 train)

University College London
Torrington Place
LONDON WC1E 7JE

For MLR and SVR, R^2 values are often negative, meaning model performance is worse than if the training test mean was used (28). When self-validated (tested on their own training set), MLR and SVR models still performed poorly, showing that the model has been 'underfit' and unable to capture the underlying patterns in the data (Appendix G). MLR's underfit is likely due to its linearity and Gaussian assumptions while the data has been shown to be non-linear and non-Gaussian (Table 7 and Figure 20) (29). Other reasons include MLR's sensitivity to outliers and the use of inter-correlated features such as Reading Week and Week of Term (30). SVR's underfit is possibly due to poorly tuned hyperparameters or inappropriate kernels, both complex variables that require thorough investigation and cross-validation (31). Underfitting could also be due to excessive regularisation, where penalties for larger feature weights are too harsh (32).

Unlike MLR and SVR, RF has high R^2 values ranging from 0.84 – 0.95, indicating that the model can identify relationships between features and occupancy well. This is likely due to its ability to model non-linear and non-Gaussian relationships, feature importance ranking, and outlier robustness (33). As an ensemble method, RF can also utilise many decision trees to reduce 'overfitting' (when a model is over-trained and is unable to generalise new data) (34).

Going forward, the RF method was used for machine learning exploration. Although tuning and scaling for SVR could be further investigated (31), RF's substantially better performance made it a more convenient method to pursue.

6.3.3 Training Set Size Investigation

After method selection, the appropriate size of training data needed to be determined. Although cross-validation is commonly used to assess performance on new data, it is not applicable to this project's occupancy data due to its temporal nature. As such, training the model on current data to predict past occupancies was not a practical investigation. Therefore, investigations were launched regarding how much time must pass until the model can accurately predict the future.

The model was trained in one-week increments, starting with just the first week of T1 until the entire T1 was trained. The remaining T1 and T2 data became the test set in each iteration,

University College London
Torrington Place
LONDON WC1E 7JE

evaluating the model's ability to predict the near and far future. As scatterplots and RSME values can be difficult to interpret, R^2 values were used as the primary measurement of accuracy, with a value of 0.85 considered here as 'accurate' (similar to the predefined '85% consistent' success criteria).

With 12 weeks in T1, 12 model iterations were tested for each area and term, resulting in 6 total investigations. Each investigation's RMSE and R^2 values are displayed graphically (Figure 23) with a highlighted region indicating when the training set includes Reading Week data. While scatterplots also show each investigation's iterative performance, only ML scatterplots are shown (Figure 24 and Figure 25) (see Appendix H for SC/SL).

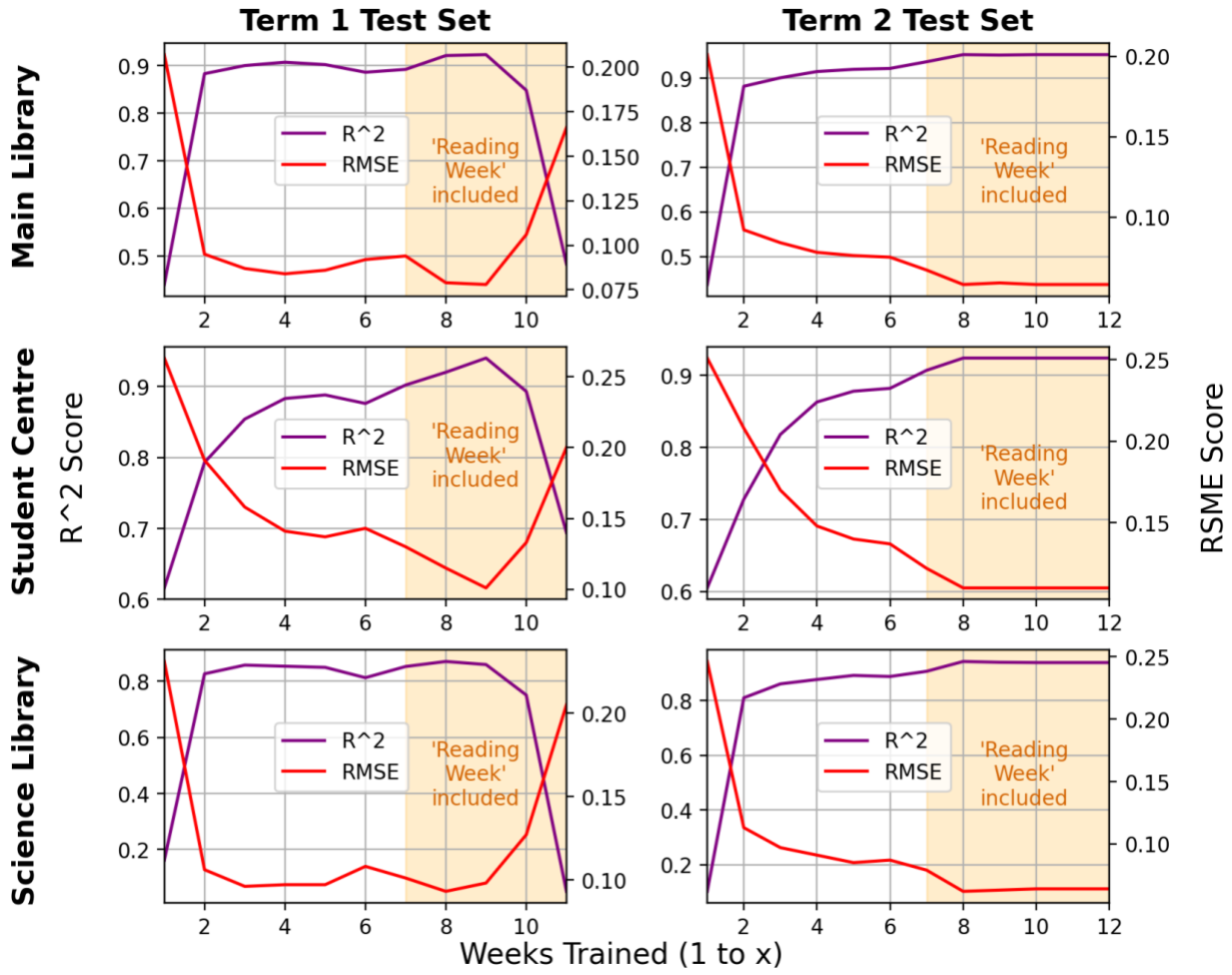

Figure 23: R^2 and RMSE scores for each investigation's incrementally trained model

University College London
Torrington Place
LONDON WC1E 7JE

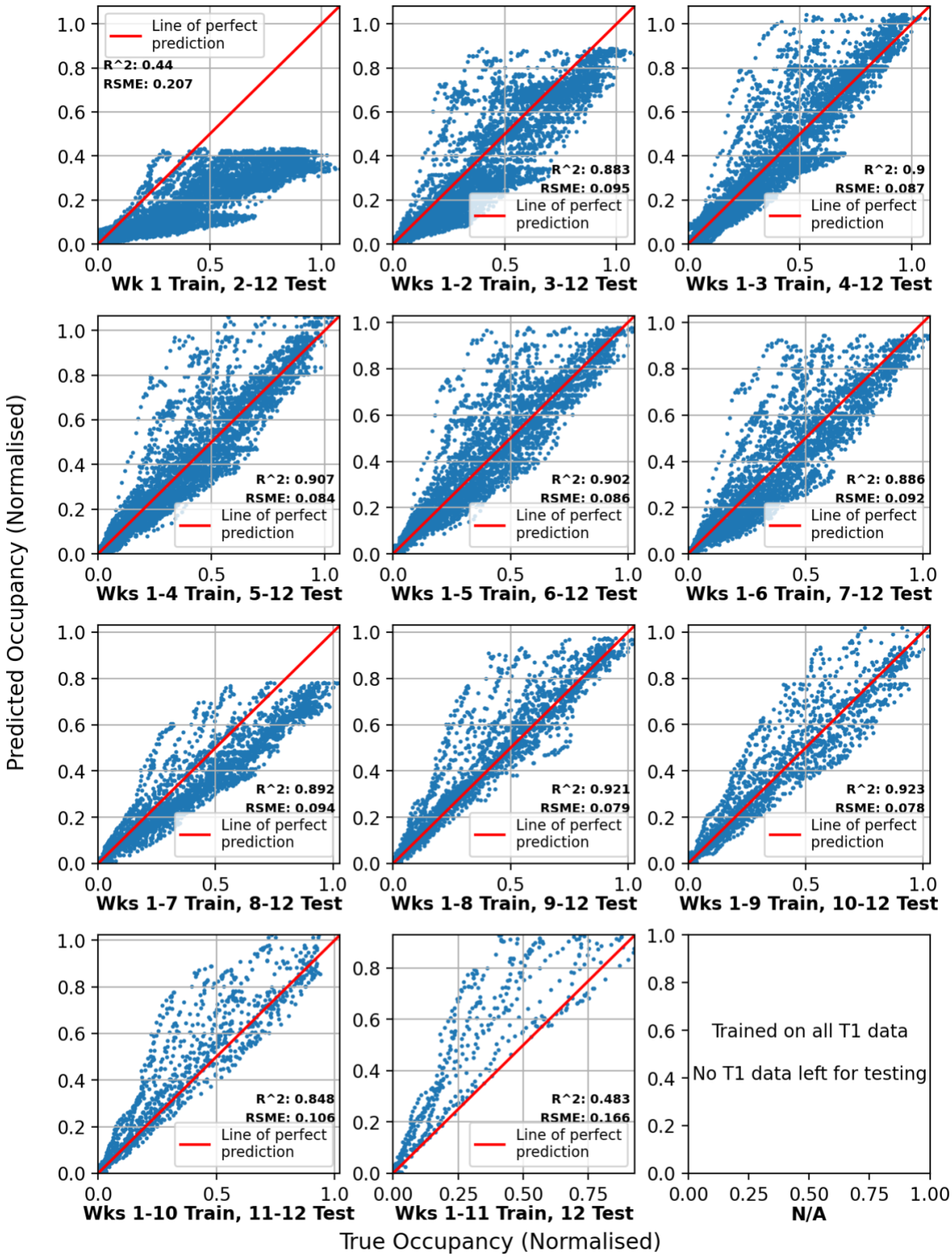

Figure 24: Predicted vs true T1 occupancy using incrementally trained RF models (ML)

University College London
Torrington Place
LONDON WC1E 7JE

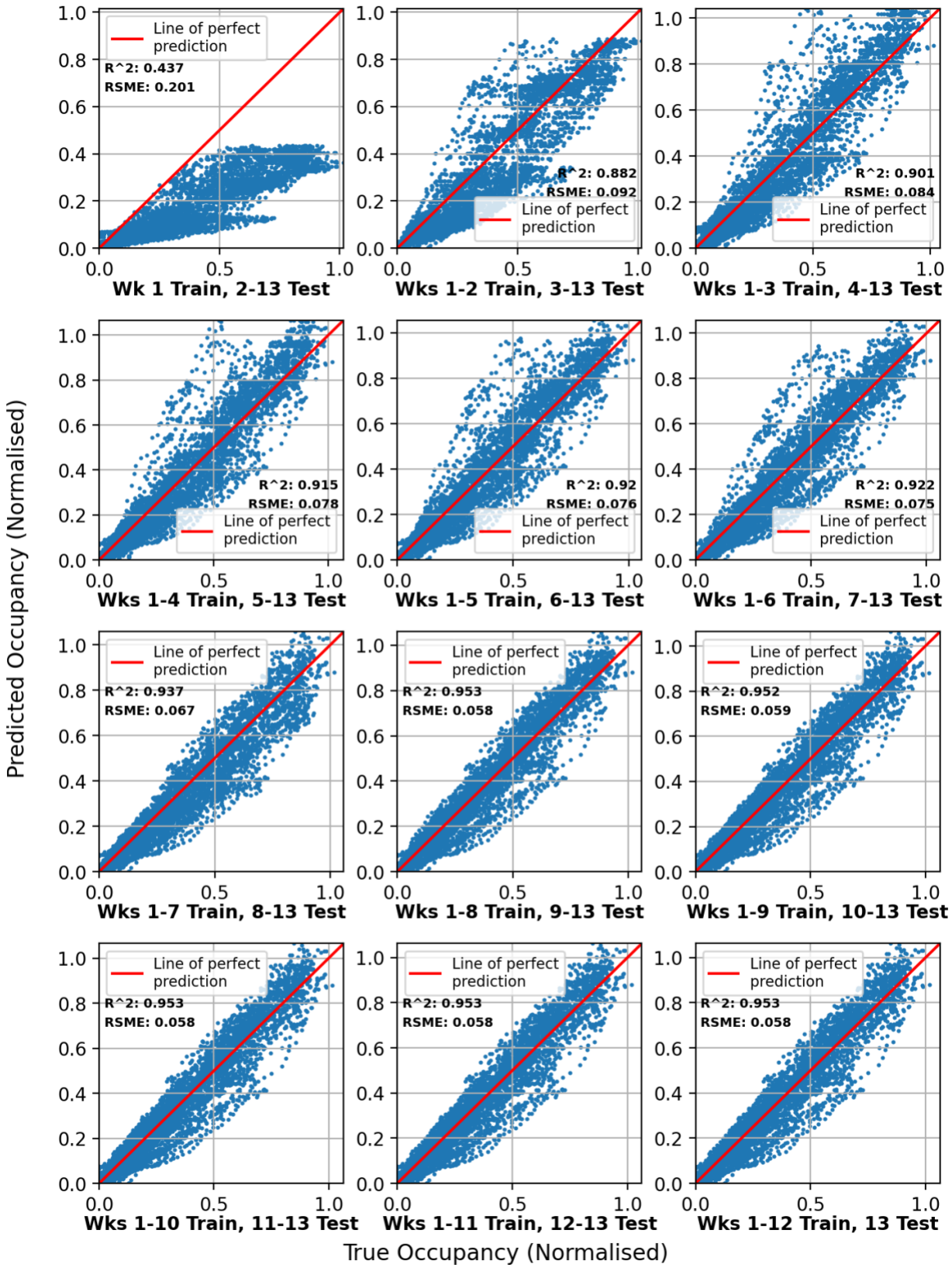

Figure 25: Predicted vs true T2 occupancy using incrementally trained RF models (ML)

University College London
Torrington Place
LONDON WC1E 7JE

As seen in Figure 23, predicting within the same term yields unexpected results as performance rapidly improves, remains strong, and then rapidly worsens after Week 9. However, this decline is only due to the iteratively decreasing test set size, making results more sensitive to remaining outliers. Most outliers are due to T1's last week (Week 12) having notably lower occupancies, being just before the winter holiday. In fact, for T1 scatterplots, these outliers cause most of the visual deviation from the LOPP. Therefore, it is easier to see the impact of the training set size when the test set size is constant (i.e., T2).

When predicting next-term occupancy, the model improves with every iteration. Notably, performance improves significantly once the test includes Week 7, able to learn from T1's Reading Week data to predict T2's Reading Week occupancy, as seen in the highlighted region of Figure 23. After Week 7, the model remains consistently accurate, having maximum R^2 scores ranging from 0.924 (SC) to 0.953 (ML). In the scatterplots, this improvement is shown by the sudden convergence of data points toward the LOPP for the Week 1-7 iteration.

In conclusion, model accuracy improves as the training set increases, especially if Reading Week is included. For predicting both same-term and next-term occupancy, the model is generally accurate within 4 weeks by R^2 standards, allowing the mobile app concept to integrate algorithm results. Further discussion and evaluation regarding model accuracy and practical application can be found in Chapter 8 –.

University College London
Torrington Place
LONDON WC1E 7JE

Chapter 7 – MOBILE APP CONCEPT

Using the research on existing mobile applications and the capabilities of this project's data analysis system, the mobile app concept includes four primary functions.

1. Historic trends
2. Real-time occupancy
3. Predicted occupancy
4. Study recommendations

The historic, real-time, and predicted functions are categorised by the area of interest with a customisable time frame filter. For study recommendations, the user only needs to specify the area, date, and time of interest to generate relevant recommendations. To help illustrate this, user flow and UX/UI concept diagrams are provided in Figure 26 and Figure 27. For preliminary concept generation examples, see Appendix I.

University College London
Torrington Place
LONDON WC1E 7JE

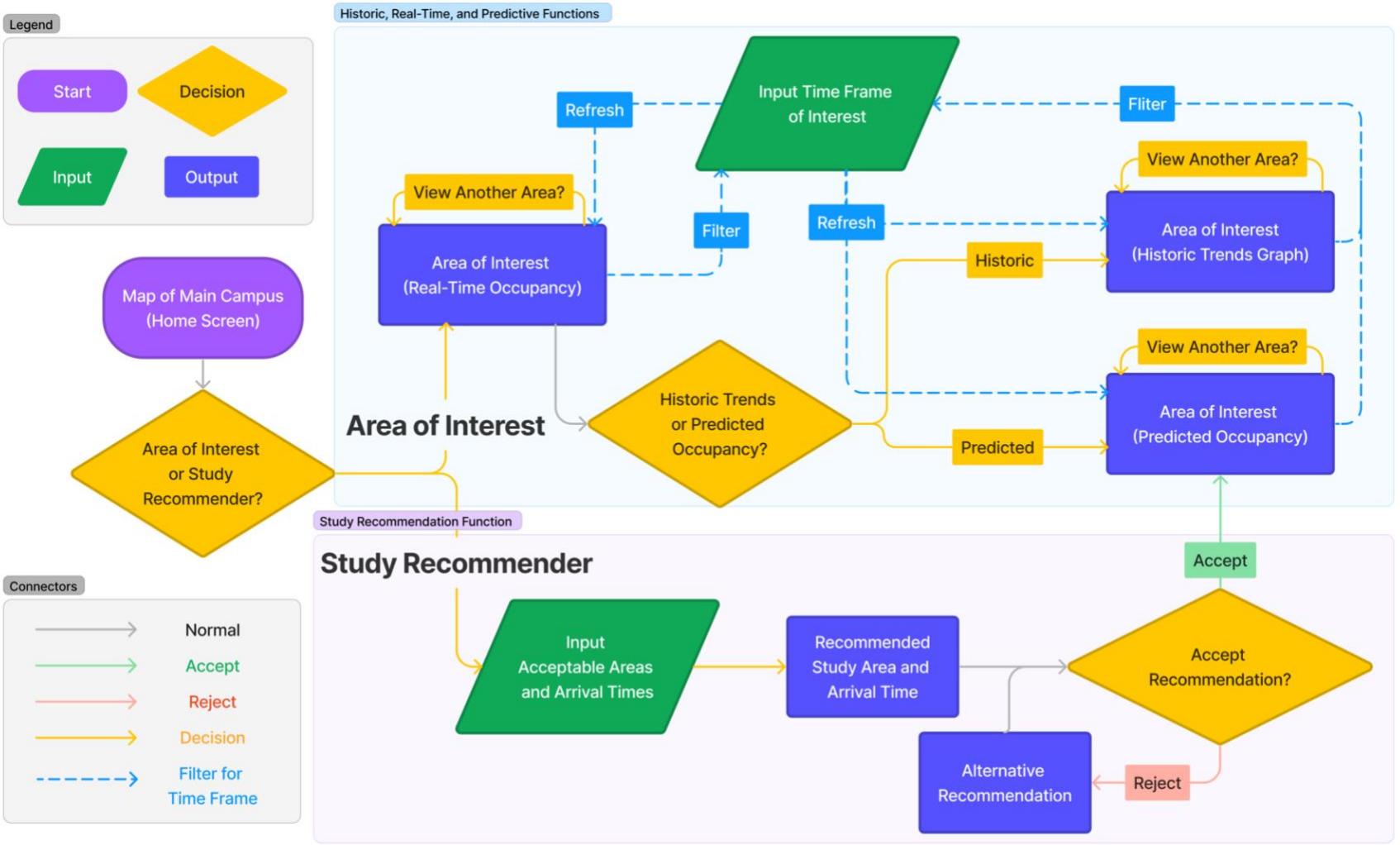

Figure 26: User flow diagram for the mobile app concept

University College London
Torrington Place
LONDON WC1E 7JE

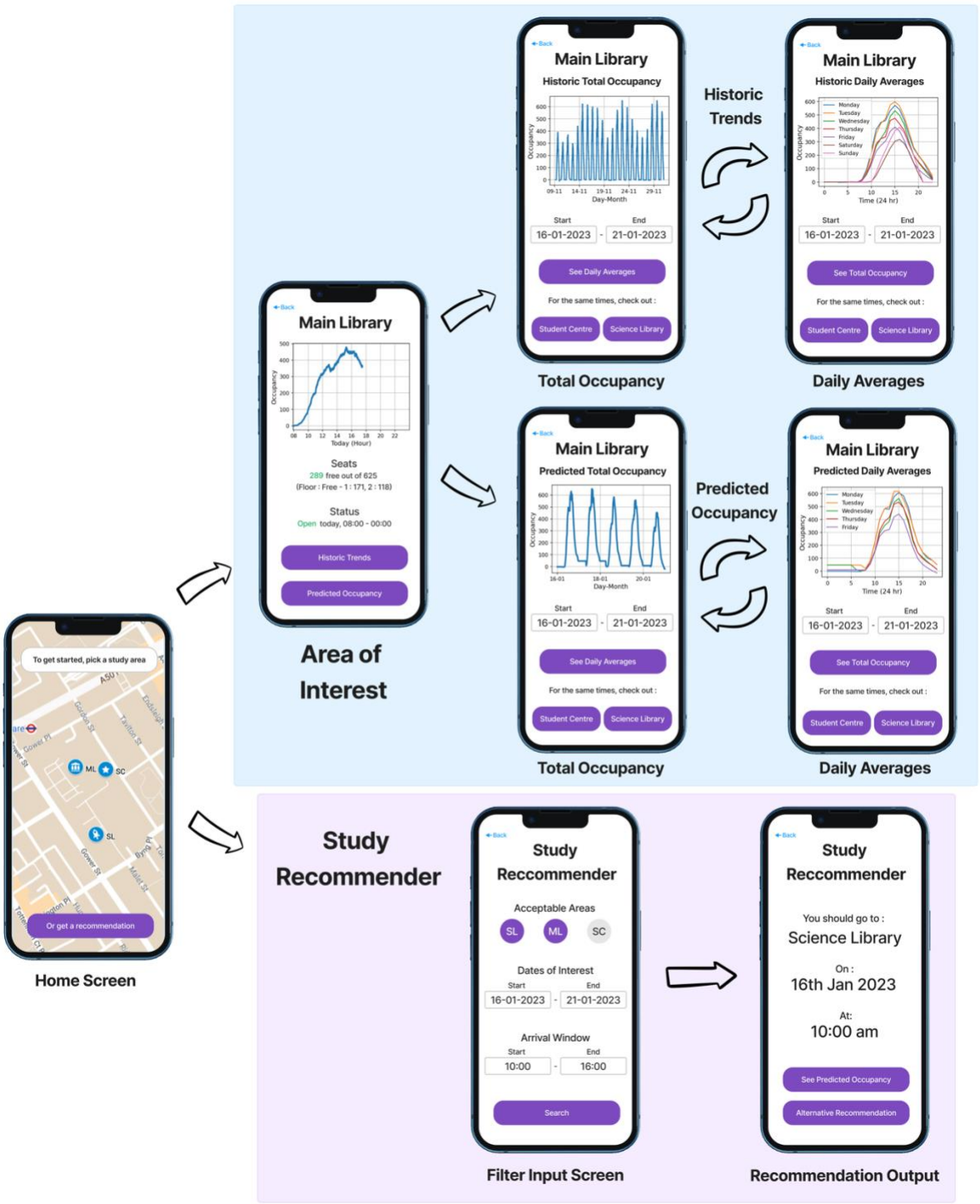

Figure 27: User flow diagram using UX/UI concept designs

University College London
Torrington Place
LONDON WC1E 7JE

7.1 Home Screen

Incorporating the iconic UCL purple, the home screen utilises a map of UCL's main campus, providing students with a distinctive and personalised way of selecting study areas and accessing historic, real-time, and predictive functions. Alternatively, the home screen also provides access to the 'study recommender' for quickly identifying free study spaces.

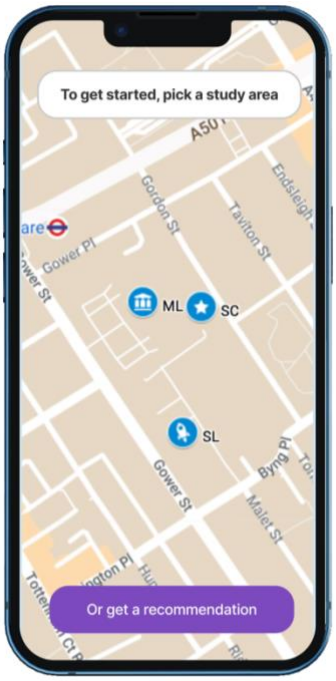

Figure 28: UX/UI concept for home screen with interactive map and study recommender access

7.2 Real-Time Occupancy

If a study area is chosen, the area's real-time occupancy is displayed, like in the UCL Go! app, in both text and graph form. As this information is publicly available, UCL's application programming interface (API) or web-scraping techniques can be used to retrieve the information for in-app display. From here, users can find out more about the area's historic or predicted

University College London
Torrington Place
LONDON WC1E 7JE

occupancies. To return to the area selection and the study recommender, a back button is placed in the upper left corner.

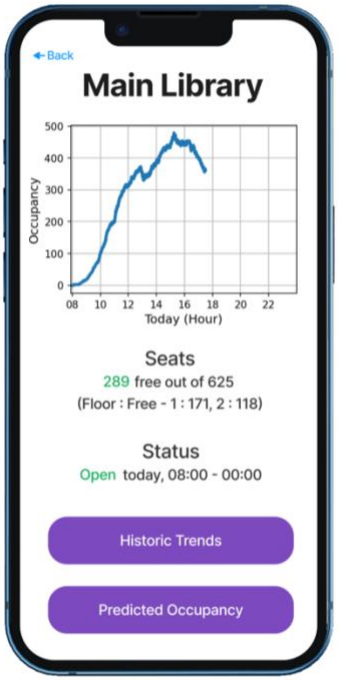

Figure 29: UX/UI concept for real-time occupancy display

7.3 Historic and Predicted Occupancy

Historic and predicted functions display either a ‘total occupancy’ or ‘daily averages’ graph according to the timeframe input. For comparison, users can switch to other areas for the same timeframe. To return to real-time occupancy, the same back button can be used.

University College London
Torrington Place
LONDON WC1E 7JE

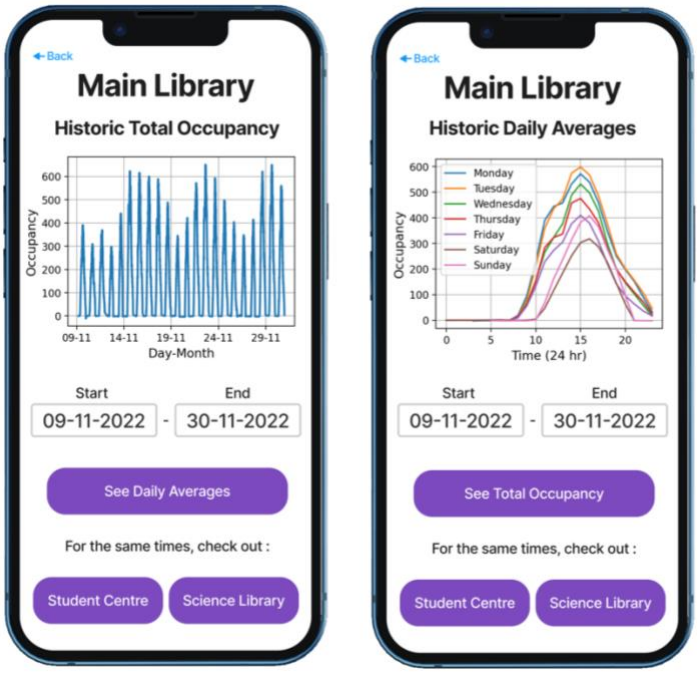

Figure 30: UX/UI concepts for historic trends

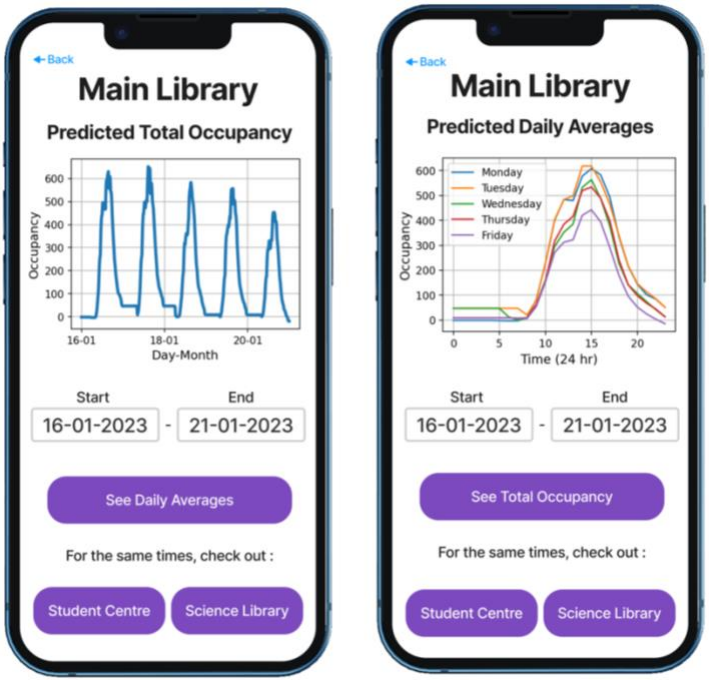

Figure 31: UX/UI concepts for predicted occupancy

University College London
Torrington Place
LONDON WC1E 7JE

7.4 Study Recommender

To use the study recommender, area, date, and arrival time filters must be applied. As study areas are categorical (nominal), the area filter is comprised of toggleable buttons, turning from grey to purple to indicate selection. For the time and date filters, only ‘start’ and ‘end’ inputs are required to establish acceptable windows.

Once applied, a recommendation is given based on *relative* crowding, specifying an area, time, and day. If accepted, the user can transition to see additional information like the predicted occupancy trend for that day. If rejected, an alternative recommendation can be requested based on the algorithm’s next-best result with the same filters. Alternatively, users can return to the input screen with the same back button.

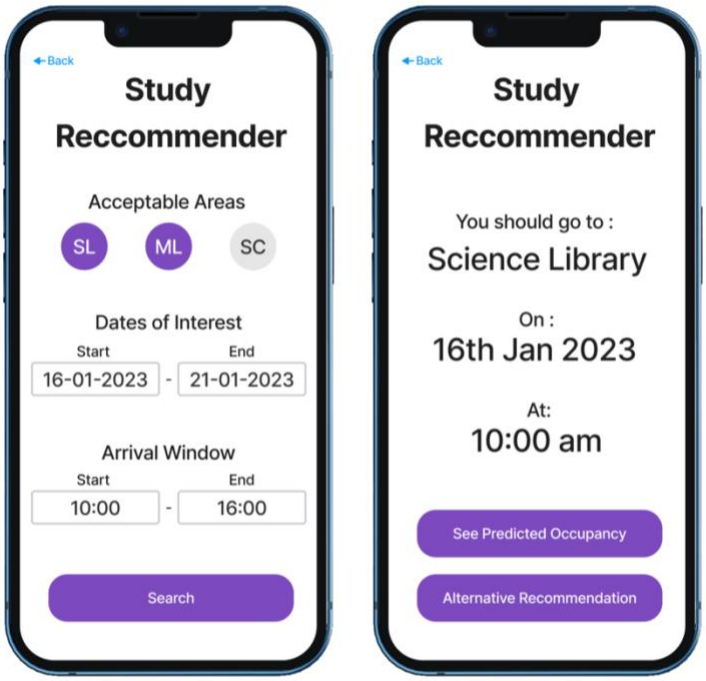

Figure 32: UX/UI concepts for study recommender input and output displays

University College London
Torrington Place
LONDON WC1E 7JE

Chapter 8 – EVALUATION AND DISCUSSION

Using the predefined design objectives, each deliverable is evaluated on its strengths, weaknesses, implications, and ambiguities.

8.1 Infrastructure Concept

8.1.1 Achieving Design Objectives

The first infrastructure-related design objective is achieved as the Gallagher SpeedStiles is an existing UCL infrastructure. The second objective is also fulfilled as the only relevant costs are related to digital infrastructure maintenance, estimated at £5k and much less than the current annual maintenance cost of £204k.

8.1.2 Weaknesses and Limitations

While an existing infrastructure, the Gallagher SpeedStiles is not optimised for people-counting, having random and systematic errors that limit data analysis accuracy. SpeedStiles also limit the number of applicable study areas, only able to monitor buildings that require ID cards upon exit. Furthermore, in buildings that accommodate hundreds of students, large spaces are reduced to a single figure, lacking the specificity that a system like the under-desk sensors can provide. Hence, while the SpeedStiles are currently more advantageous, the under-desk sensors present a higher potential for future application.

University College London
Torrington Place
LONDON WC1E 7JE

8.2 Data Analysis System

8.2.1 Achieving Design Objectives

To achieve autonomy, the system is comprised of a series of procedural Python scripts, from extracting historic trends to predicting future occupancy. While random errors still require manual correction, all scripts are fully automatic, only requiring custom inputs such as time filters.

To evaluate the accuracy and consistency of the prediction algorithm, the predefined success criteria can be applied, as seen in Figures 33 – 35 (See Appendix J for SC/SL scatterplots).

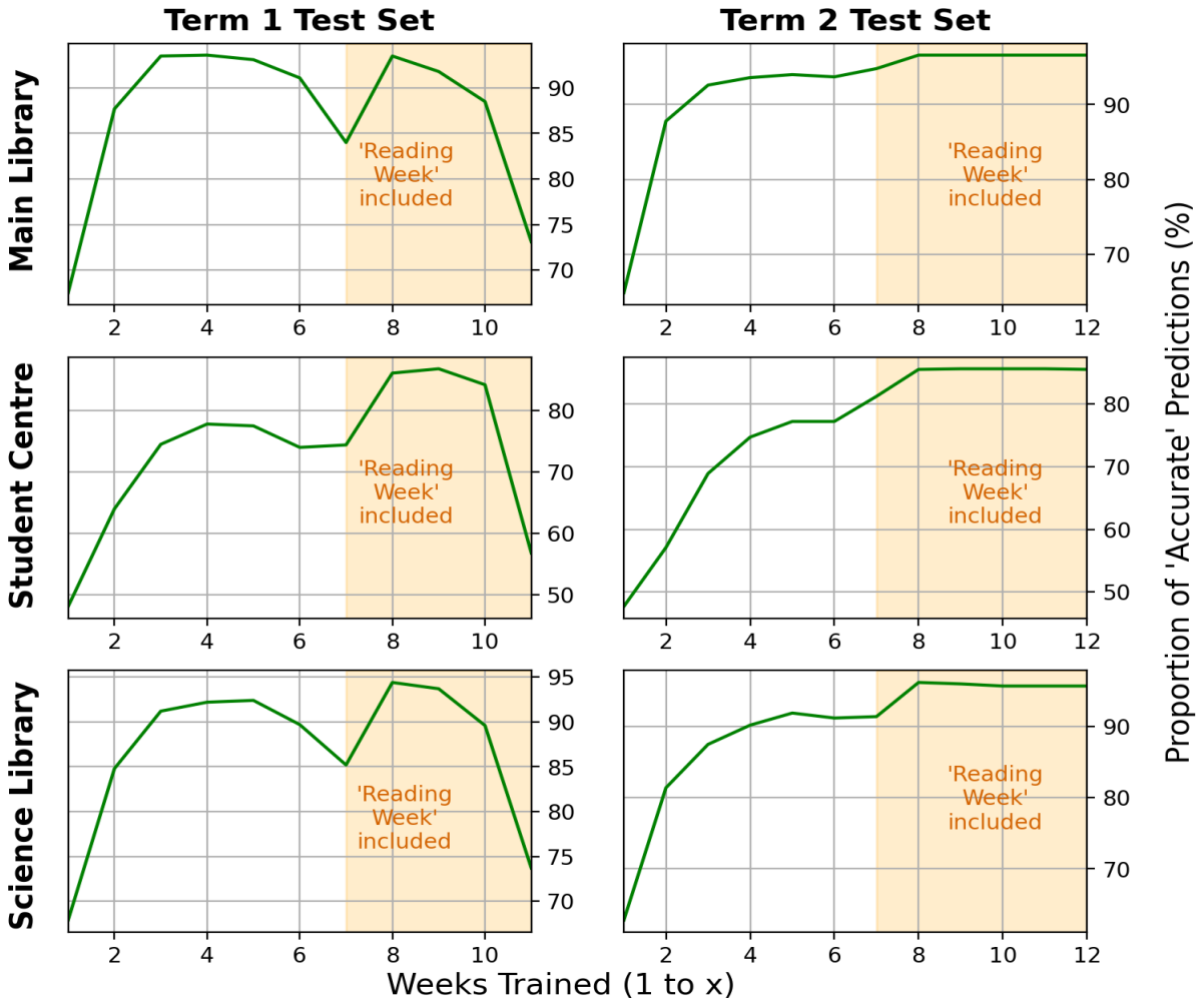

Figure 33: Proportion of 'accurate' predictions for each investigation's incrementally trained model

University College London
Torrington Place
LONDON WC1E 7JE

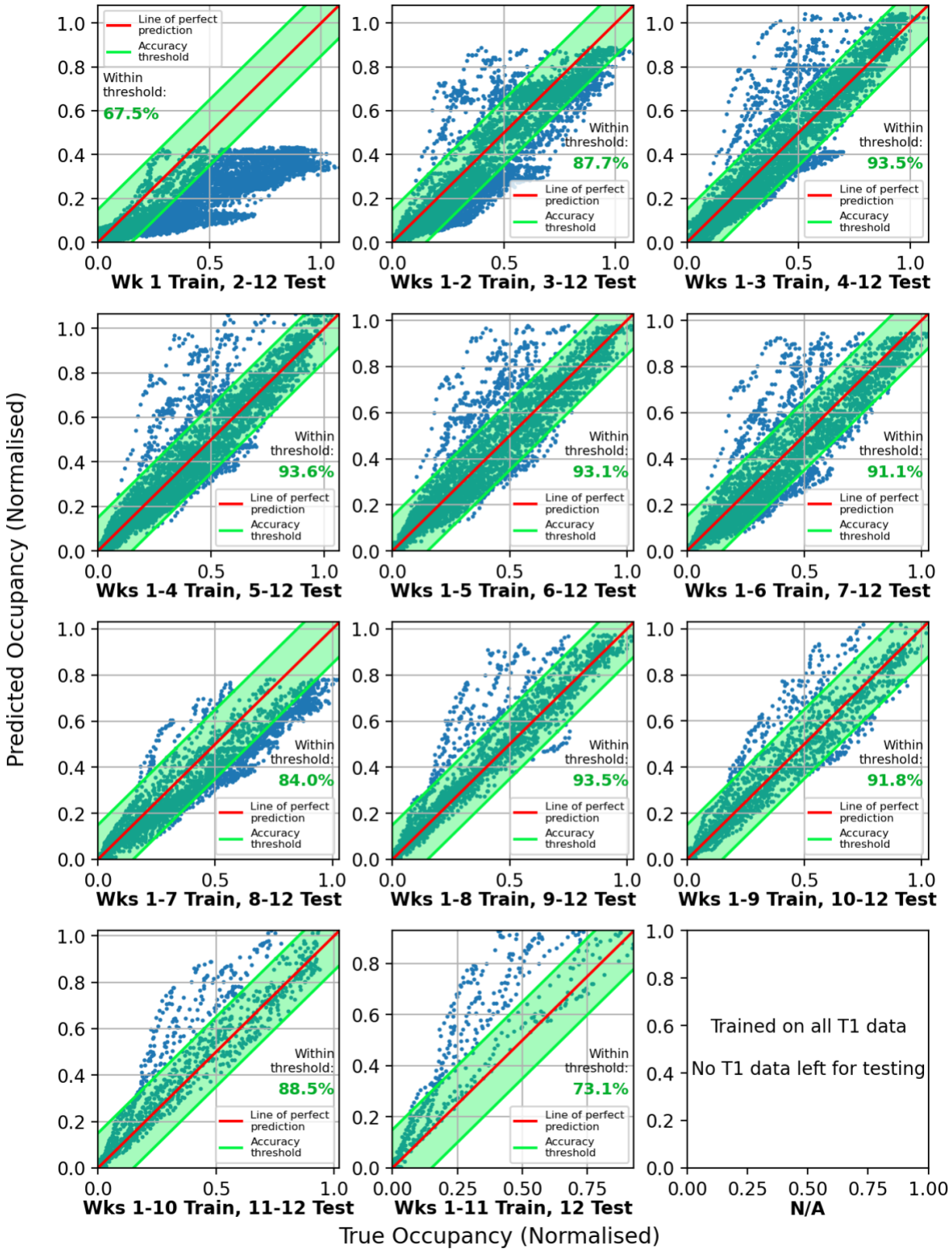

Figure 34: Predicted vs true T1 occupancy using incrementally trained RF models (accuracy threshold) (ML)

University College London
Torrington Place
LONDON WC1E 7JE

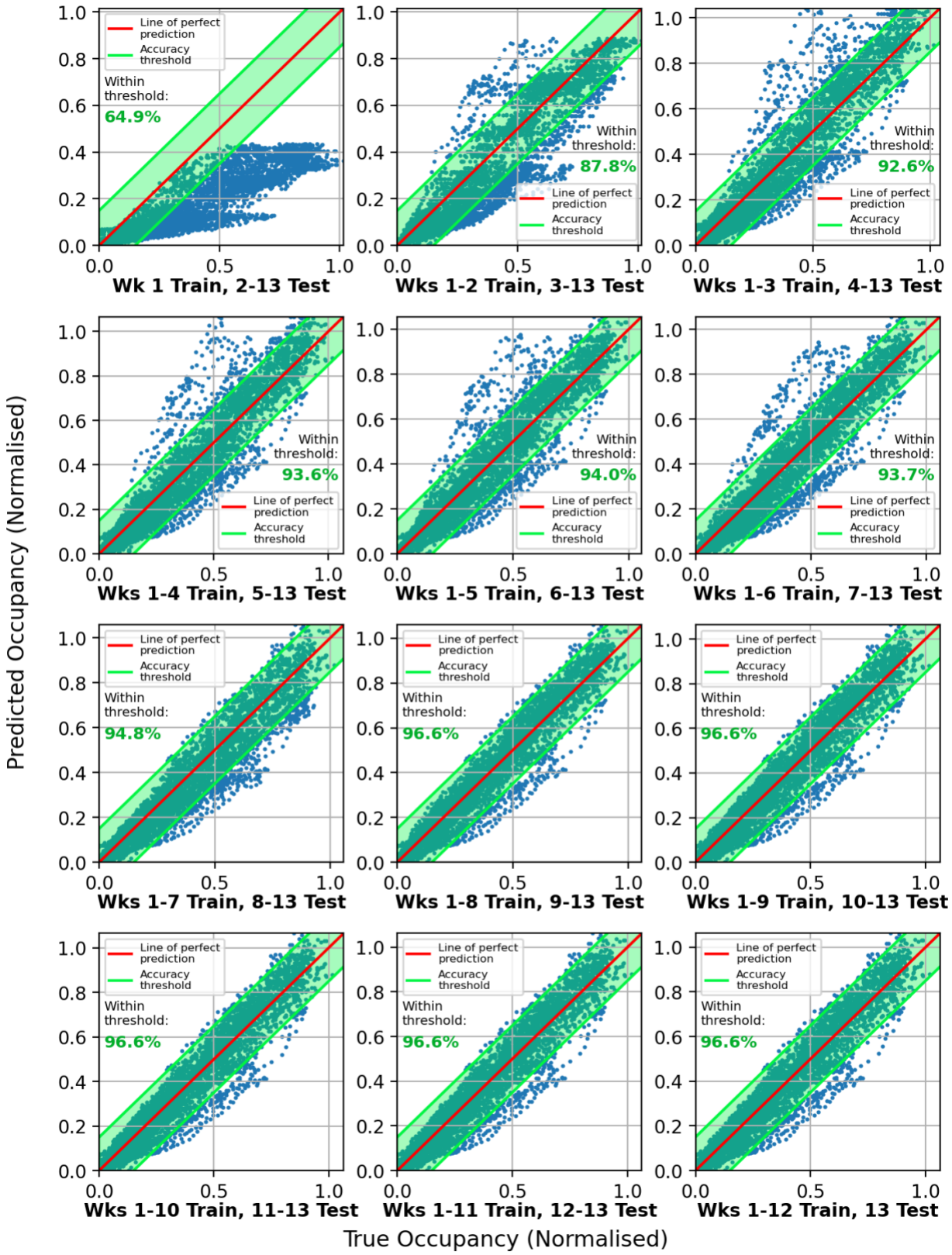

Figure 35: Predicted vs true T2 occupancy using incrementally trained RF models (accuracy threshold) (ML)

University College London
Torrington Place
LONDON WC1E 7JE

Using the accuracy success criteria, an 'accuracy threshold' (green) can be highlighted in Figure 34 and Figure 35 (see Appendix J for SC/SL scatterplots). The proportion of points that fall within the threshold is the model's 'consistency,' which can then be plotted as a line graph (Figure 33).

As seen in Figure 33, the model eventually reaches 85% 'consistency' in all 6 investigations. For the ML and SL, the model achieves this standard by Week 3 before improving to approximately 95%. For the SC, the standard is not achieved until Reading Week data has been trained at Week 7 or 8. In fact, as with the RMSE and R^2 metrics, consistency scores improve greatly after Reading Week but eventually worsen in T1 due to the decreasing test size.

8.2.2 Weaknesses and Limitations

As seen in Figure 33, the model most struggles with SC predictions, improving slower than other areas through each iteration and never achieving 90% consistency. This is due to the SC's relative popularity and availability, attracting more students and remaining open when other areas close.

Another limitation is that only 2022-2023 data is analysed, being the first fully in-person academic year since the COVID-19 pandemic. In subsequent years, the prediction model would have access to previous annual data, possibly making previous performance evaluations too harsh.

While model accuracy is achieved within T1, it is arguable that an entire term is necessary to build the prediction model. Unless features such as the holiday proximity or Reading Weeks are included, predictions for some weeks can be much less accurate than for the overall test set. In Figure 36, the iterative model's ML predictions for only the subsequent week in T1 are displayed, showing that the model performs poorly for specifically Weeks 7 and 12 (Reading Week and the last week of T1) (See for Appendix K SC/SL).

University College London
Torrington Place
LONDON WC1E 7JE

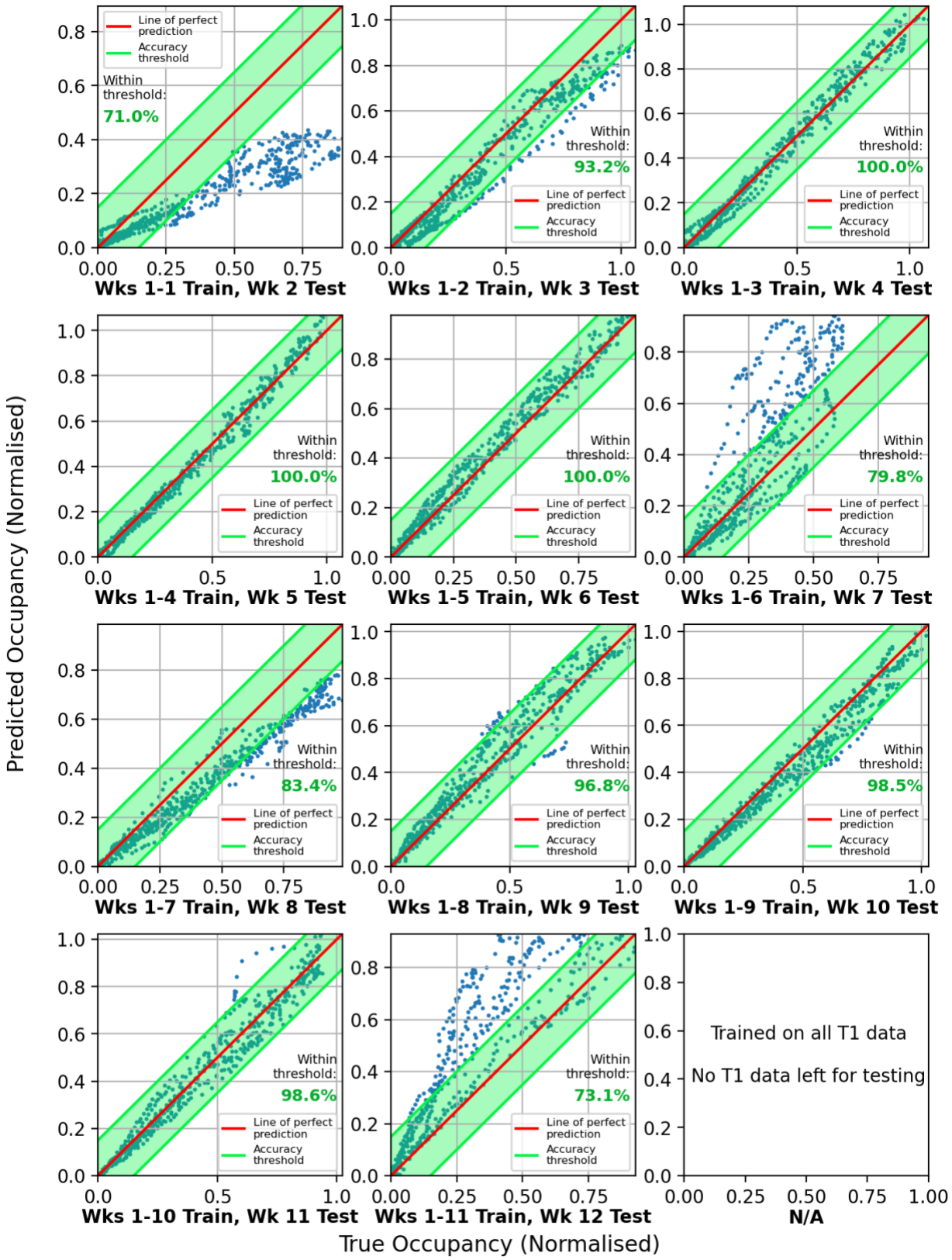

Figure 36: Predicted vs true T1 occupancy using incrementally trained RF models (1-week test sets) (ML)

University College London
Torrington Place
LONDON WC1E 7JE

8.3 Mobile App Concept

8.3.1 Achieving Design Objectives

Using the pre-defined success criteria, all app-related design objectives are achieved. Firstly, real-time occupancies (available through UCL Go!) along with historic, predictive, and study recommender functions (from the data analysis system) are integrated into the app concept. Secondly, for UX/UI features, large and minimal text is used along with a contrasting two-tone colour palette and simple area/time filters.

8.3.2 Weaknesses and Limitations

Currently, the mobile app concept focuses on UX/UI design and does not encompass app-engineering concepts such as security or log-in features that apps like UCL Go! require. Furthermore, this concept has not been user-tested and may not accurately reflect the needs of a typical student, requiring consultation with UX/UI designers and app developers.

8.4 Project Approach

In addition to the three deliverables, the project’s double-diamond framework and technical approach are also evaluated.

8.4.1 Double-Diamond Framework

Using the double-diamond framework resulted in several benefits, including a large focus on the problem context and user requirements. Through the 'Discover' and 'Define' stages, UCL's specific problems involving data accessibility and student satisfaction as well as the existing infrastructure were identified. This information was critical in establishing a basis for the practical implementation of the project and understanding user needs.

University College London
Torrington Place
LONDON WC1E 7JE

While high-quality problem definition and research processes were ensured, these processes were incredibly time-consuming, resulting in arguably unnecessary investigations such as non-UCL people-counting sensor systems. As a holistic approach, the double-diamond could have been replaced by more specific and efficient methods such as an agile methodology or design sprints. Specifically beneficial in software development projects like this one, both methods offer a more flexible and iterative approach, often involving rapid prototyping.

8.4.2 Technical Approach

The project's technical approach was well-executed, involving sequential analysis of datasets to extract important features and observations, followed by the selection of an appropriate machine learning method and train-test split. This aligns with the 'Develop' and 'Deliver' stages of the double-diamond framework, using data exploration to deliver an accurate data analysis system.

However, it is arguable that SVR was not thoroughly explored during method selection. Having more hyperparameters than RF, SVR's complexity likely led to its poor performance and could have benefited from feature scaling and proper tuning.

Furthermore, the incompleteness of T2 data did not allow the model to be tested on T2's last weeks which likely contained outliers with the spring holiday approaching. Hence, the model may have performed better with T2 predictions due to only very predictable weeks included in the test set.

While the technical investigation was thorough, the narrow focus given to machine learning as a prediction method is also flawed. To find the best predictive algorithm, popular statistical methods like AutoRegressive Integrated Moving Average (ARIMA) models should have also been investigated. Capable of capturing underlying patterns and accommodating irregular fluctuations, an ARIMA model may prove to be better at predicting occupancy (35).

University College London
Torrington Place
LONDON WC1E 7JE

Chapter 9 – CONCLUSION

The data analysis system accurately predicted occupancies of all areas, achieving predefined success criteria with as little as 3 weeks of training data. Concurrently, the infrastructure concept successfully incorporated existing UCL systems with minimal incremental costs (£5000 annually) and the mobile app concept illustrated user procedures via user-flow and concept diagrams. However, limitations include limited accuracy regarding SC predictions, requiring at least 7 weeks of training data due to its high popularity and unrestricted opening hours. Furthermore, the infrastructure and app concepts remain nascent and unproven, requiring extensive development before real-world implementation can commence.

Project implications involve improving occupancy analysis at UCL and beyond, being applicable to any organisation that utilises electronic turnstiles. As a predominantly digital solution, such organisations could avoid investments in expensive commercial sensors while benefitting from advanced occupancy analysis and associated applications. These applications include optimising opening hours, floorplan designs, and space allocation.

9.1 Future Work

As occupancy analysis is a vast field, this project has many avenues for further exploration.

9.1.1 Occupancy Analysis

Firstly, larger timescales should be applied to evaluate annual prediction performance and explore features such as exam proximity and total student enrolment. Secondly, ARIMA models should also be explored for comparison against machine learning techniques in terms of performance and simplicity.

University College London
Torrington Place
LONDON WC1E 7JE

9.1.2 Practical Application

For alternative IoT systems, under-desk sensors should be pursued, allowing for more accurate data analysis and an increase in study space coverage. Once physically established, cloud-computing trials should support prototype app development and allow for small-scale user trials. These trials can be leveraged to demonstrate the solution's practical utility while more accurately defining user requirements (i.e., most popular area/time filters).

Furthermore, the model's predictive capabilities could be applied to new areas such as UCL East or early architectural designs. Using predicted occupancies, administrators and architects could optimise opening hours and floorplans to reduce energy consumption and plan efficient evacuation routes.

University College London
Torrington Place
LONDON WC1E 7JE

REFERENCES

1. University College London. Student Numbers by Method of Study 2012-2022 [Internet]. London; 2022 [cited 2023 Feb 11]. Available from: https://www.ucl.ac.uk/srs/student-statistics.

2. Wang T. Learning to Do More With Less: Adapting Campus Security for Lean Times. asmag [Internet]. 2012 [cited 2022 Oct 22]. Available from: https://www.asmag.com/showpost/12628.aspx?mv=&pages=2.

3. Cetinkaya HH, Akcay M. People Counting at Campuses. Procedia Soc Behav Sci. Elsevier BV; 2015; 182:732–6.

4. Monti L, Mirri S, Prandi C, Salomoni P. Smart Sensing Supporting Energy-Efficient Buildings: On Comparing Prototypes for People Counting. Proceedings of the 5th EAI International Conference on Smart Objects and Technologies for Social Good [Internet]. New York, NY, USA: ACM; 2019 [cited 2022 Oct 22]. Available from: https://doi.org/10.1145/3342428.3342691.

5. University of Edinburgh. Space occupancy monitoring | The University of Edinburgh [Internet]. 2020 [cited 2022 Oct 24]. Available from: https://www.ed.ac.uk/information-services/students/study-space/space-occupancy-monitoring-pilot.

6. Sutjarittham T, Habibi Gharakheili H, Kanhere SS, Sivaraman V. Experiences with IoT and AI in a Smart Campus for Optimizing Classroom Usage. IEEE Internet Things J. Institute of Electrical and Electronics Engineers Inc.; 2019; 6(5):7595–607.

7. European Commission. Space and the City. European Commission [Internet]. 2023 [cited 2023 Jan 19]. Available from: https://urban.jrc.ec.europa.eu/thefutureofcities/space-and-the-city#the-chapter.

8. Office of Space Optimisation. The Importance of Space. University of Colorado, Boulder [Internet]. 2023 [cited 2023 Jan 19]. Available from: https://www.colorado.edu/space-optimization/governance/history/space-utilization-and-optimization-initiative.

University College London
Torrington Place
LONDON WC1E 7JE

9. Furini M, Mandreoli F, Martoglia R, Montangero M. IoT: Science fiction or real revolution? Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST [Internet]. Springer Verlag; 2017 [cited 2022 Oct 22]; 195 LNICST:96–105. Available from: https://link.springer.com/chapter/10.1007/978-3-319-61949-1_11.

10. Greco A, Saggese A, Vento B. A Robust and Efficient Overhead People Counting System for Retail Applications. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) [Internet]. Springer Science and Business Media Deutschland GmbH; 2022 [cited 2022 Oct 22]; 13232 LNCS:139–50. Available from: https://link.springer.com/chapter/10.1007/978-3-031-06430-2_12.

11. Stewart M. The Actual Difference Between Statistics and Machine Learning | by Matthew Stewart, PhD | Towards Data Science. Towards Data Science [Internet]. 2019 [cited 2023 Mar 5]. Available from: https://towardsdatascience.com/the-actual-difference-between-statistics-and-machine-learning-64b49f07ea3.

12. Bell J. Machine Learning and the City. Machine Learning and the City. John Wiley & Sons; 2022.

13. Breiman L. Random forests. Mach Learn [Internet]. Springer; 2001 [cited 2023 Feb 26]; 45(1):5–32. Available from: https://link.springer.com/article/10.1023/A:1010933404324.

14. Cortes C, Vapnik V, Saitta L. Support-vector networks. Machine Learning 1995 20:3 [Internet]. Springer; 1995 [cited 2023 Feb 26]; 20(3):273–97. Available from: https://link.springer.com/article/10.1007/BF00994018.

15. Silverman J. NU administration removes occupancy sensors in ISEC in response to privacy, ethical concerns - The Huntington News [Internet]. 2022 [cited 2023 Feb 15]. Available from: https://huntnewsnu.com/69260/campus/nu-administration-removes-occupancy-sensors-in-isec-in-response-to-privacy-ethical-concerns/.

16. White JP, Dennis S, Tomko M, Bell J, Winter S. Paths to social licence for tracking-data analytics in university research and services. PLoS One [Internet]. Public Library of

University College London
Torrington Place
LONDON WC1E 7JE

Science; 2021 [cited 2023 Feb 15]; 16(5):e0251964. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0251964.

17. Tchounwou B, Herr RM, Tunahan GI, Altamirano H. Seating Behaviour of Students before and after the COVID-19 Pandemic: Findings from Occupancy Monitoring with PIR Sensors at the UCL Bartlett Library [Internet]. 2022 [cited 2023 Feb 13]. Available from: https://doi.org/10.3390/ijerph192013255.

18. Dawe M. Electronic Access Control Turnstile Integration Specification Guidance Document Specification for the installation of turnstiles, and their integration with UCL's Gallagher access control system Contents: 1.0 General requirements [Internet]. 2017 [cited 2023 Feb 13]. Available from: https://www.ucl.ac.uk/estates/sites/estates/files/turnstile_integration_specification.pdf.

19. University College London. Attendance at UCL | Students - UCL – University College London. University College London [Internet]. 2023 [cited 2023 Feb 13]. Available from: https://www.ucl.ac.uk/students/life-ucl/attendance-ucl.

20. Terabee. Terabee People Counting L-XL - Improve facility efficiency with accurate people counting data [Internet]. 2023 [cited 2023 Mar 28]. Available from: https://www.terabee.com/shop/people-counting/terabee-people-counting-l/.

21. Occuspace Inc. Waitz University on the App Store [Internet]. 2023 [cited 2023 Mar 29]. Available from: https://apps.apple.com/us/app/waitz-university/id1346827447.

22. Amazon Web Services. How to Build a Mobile App - Amazon Web Services (AWS) [Internet]. 2023 [cited 2023 Mar 29]. Available from: https://aws.amazon.com/startups/start-building/how-to-build-a-mobile-app/.

23. Google Firebase. Firebase Pricing [Internet]. 2023 [cited 2023 Mar 29]. Available from: https://firebase.google.com/pricing#blaze-calculator.

24. Georgiou M. Mobile App Maintenance: Importance, Types & Cost (in 2023) [Internet]. 2023 [cited 2023 Mar 29]. Available from: https://imaginovation.net/blog/importance-mobile-app-maintenance-cost/#.

University College London
Torrington Place
LONDON WC1E 7JE

25. Georgiou M. Mobile App Maintenance: Importance, Types & Cost (in 2023) [Internet]. 2023 [cited 2023 Apr 22]. Available from: https://imaginovation.net/blog/importance-mobile-app-maintenance-cost/.

26. University College London. Opening hours | Library Services - UCL – University College London [Internet]. 2023 [cited 2023 Mar 29]. Available from: https://www.ucl.ac.uk/library/using-library/opening-hours.

27. Mohd Razali N, Bee Wah Y. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. Journal of Statistical Modeling and Analytics. 2011; 2(1):13–4.

28. Chicco D, Warrens MJ, Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput Sci [Internet]. PeerJ Inc.; 2021 [cited 2023 Apr 6]; 7:1–24. Available from: https://peerj.com/articles/cs-623.

29. Tranmer M, Murphy J, Elliot M, Pampaka M. Multiple Linear Regression (2 nd Edition) [Internet]. 2020 [cited 2023 Apr 6]. Available from: https://hummedia.manchester.ac.uk/institutes/cmist/a.

30. Uyanık GK, Güler N. A Study on Multiple Linear Regression Analysis. Procedia Soc Behav Sci. Elsevier; 2013; 106:234–40.

31. Suthaharan S. Support Vector Machine [Internet]. Springer, Boston, MA; 2016 [cited 2023 Apr 6]; 207–35. Available from: https://link.springer.com/chapter/10.1007/978-1-4899-7641-3_9.

32. Xu XUHUAN H, Caramanis C, Mannor S, Smola A. Robustness and Regularization of Support Vector Machines. Journal of Machine Learning Research. 2009; 10:1485–510.

33. Biau G, Scornet E. A random forest guided tour. Test [Internet]. Springer New York LLC; 2016 [cited 2023 Apr 6]; 25(2):197–227. Available from: https://link.springer.com/article/10.1007/s11749-016-0481-7.

University College London
Torrington Place
LONDON WC1E 7JE

34. Ao Y, Li H, Zhu L, Ali S, Yang Z. The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling. J Pet Sci Eng. Elsevier; 2019; 174:776–89.

35. Shumway RH, Stoffer DS. ARIMA Models [Internet]. Springer, Cham; 2017 [cited 2023 Apr 11]; 75–163. Available from: https://link.springer.com/chapter/10.1007/978-3-319-52452-8_3.

University College London
Torrington Place
LONDON WC1E 7JE

APPENDICES

Appendix A

ACADEMIC YEAR 2022/2023

TERM 1 MON 26 SEP 2022 - FRI 16 DEC 2022
TERM 2 MON 09 JAN 2023 - FRI 24 MAR 2023
TERM 3 MON 24 APR 2023 - FRI 09 JUN 2023

UCL CLOSES FRI 23 DEC 2022 - MON 2 JAN 2023
UCL CLOSES WED 5 APR 2023 - WED 12 APR 2023
READING WEEKS W/C MON 7 NOV 2022 & MON 13 FEB 2023
BANK HOLIDAYS MON 29 AUG 2022, MON 01 MAY & 29 MAY 2023

AUGUST

WK	M	T	W	T	F	S	S
1	29	30	31				

SEPTEMBER

WK	M	T	W	T	F	S	S
1				1	2	3	4
2	5	6	7	8	9	10	11
3	12	13	14	15	16	17	18
4	19	20	21	22	23	24	25
5	26	27	28	29	30		

OCTOBER

WK	M	T	W	T	F	S	S
5						1	2
6	3	4	5	6	7	8	9
7	10	11	12	13	14	15	16
8	17	18	19	20	21	22	23
9	24	25	26	27	28	29	30
10	31						

NOVEMBER

WK	M	T	W	T	F	S	S
10		1	2	3	4	5	6
11	7	8	9	10	11	12	13
12	14	15	16	17	18	19	20
13	21	22	23	24	25	26	27
14	28	29	30				

DECEMBER

WK	M	T	W	T	F	S	S
14				1	2	3	4
15	5	6	7	8	9	10	11
16	12	13	14	15	16	17	18
17	19	20	21	22	23	24	25
18	26	27	28	29	30	31	

JANUARY

WK	M	T	W	T	F	S	S
18							1
19	2	3	4	5	6	7	8
20	9	10	11	12	13	14	15
21	16	17	18	19	20	21	22
22	23	24	25	26	27	28	29
23	30	31					

FEBRUARY

WK	M	T	W	T	F	S	S
23			1	2	3	4	5
24	6	7	8	9	10	11	12
25	13	14	15	16	17	18	19
26	20	21	22	23	24	25	26
27	27	28					

MARCH

WK	M	T	W	T	F	S	S
27			1	2	3	4	5
28	6	7	8	9	10	11	12
29	13	14	15	16	17	18	19
30	20	21	22	23	24	25	26
31	27	28	29	30	31		

APRIL

WK	M	T	W	T	F	S	S
31						1	2
32	3	4	5	6	7	8	9
33	10	11	12	13	14	15	16
34	17	18	19	20	21	22	23
35	24	25	26	27	28	29	30

MAY

WK	M	T	W	T	F	S	S
36	1	2	3	4	5	6	7
37	8	9	10	11	12	13	14
38	15	16	17	18	19	20	21
39	22	23	24	25	26	27	28
40	29	30	31				

JUNE

WK	M	T	W	T	F	S	S
40				1	2	3	4
41	5	6	7	8	9	10	11
42	12	13	14	15	16	17	18
43	19	20	21	22	23	24	25
44	26	27	28	29	30		

JULY

WK	M	T	W	T	F	S	S
44						1	2
45	3	4	5	6	7	8	9
46	10	11	12	13	14	15	16
47	17	18	19	20	21	22	23
48	24	25	26	27	28	29	30
49	31						

AUGUST

WK	M	T	W	T	F	S	S
49		1	2	3	4	5	6
50	7	8	9	10	11	12	13
51	14	15	16	17	18	19	20
52	21	22	23	24	25	26	27

Figure A1: UCL academic calendar for 2022-2023

University College London
Torrington Place
LONDON WC1E 7JE

Appendix B

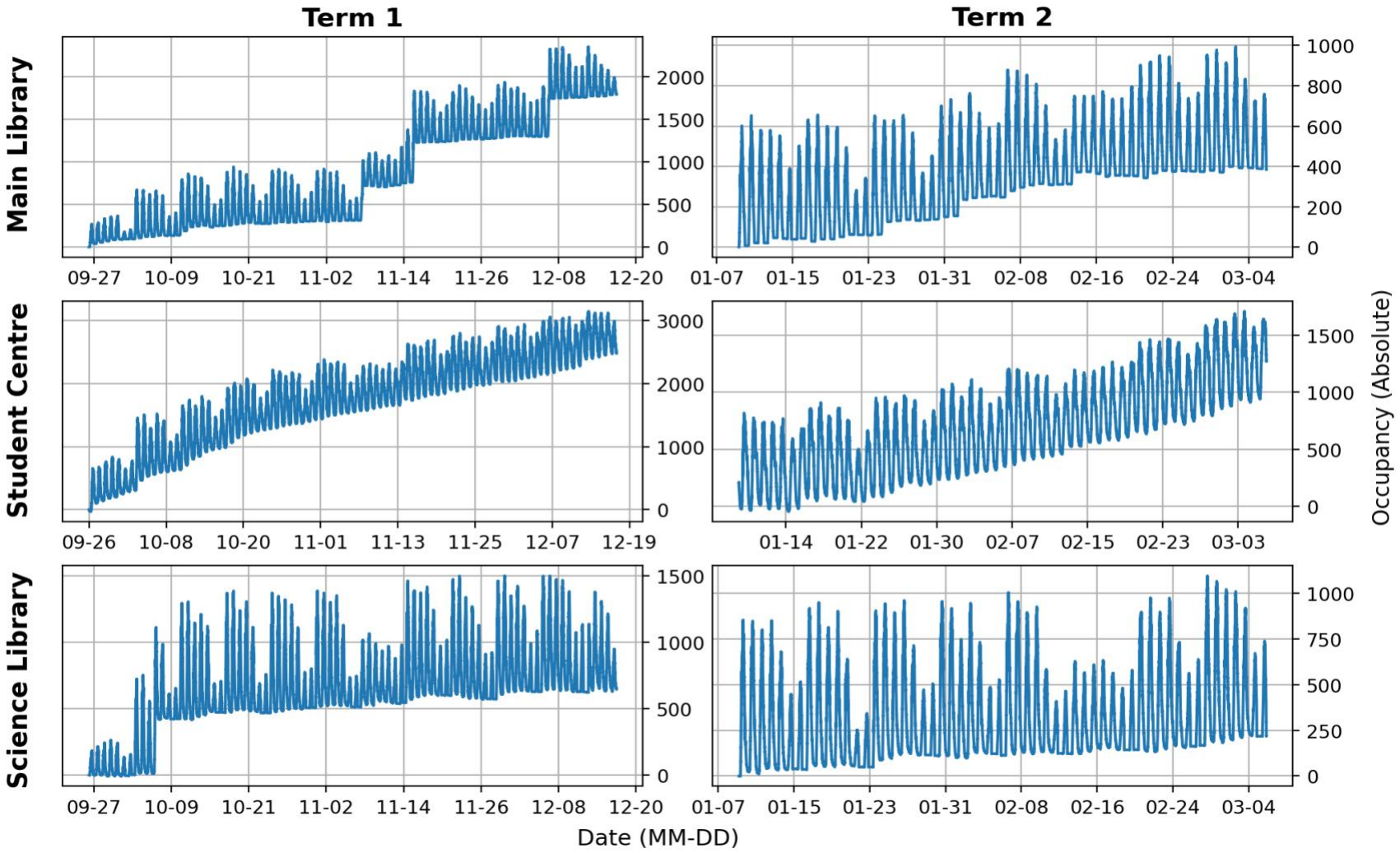

Figure A2: Uncorrected ML, SC, and SL occupancies for T1 and T2 (random and systematic error)

University College London
Torrington Place
LONDON WC1E 7JE

Appendix C

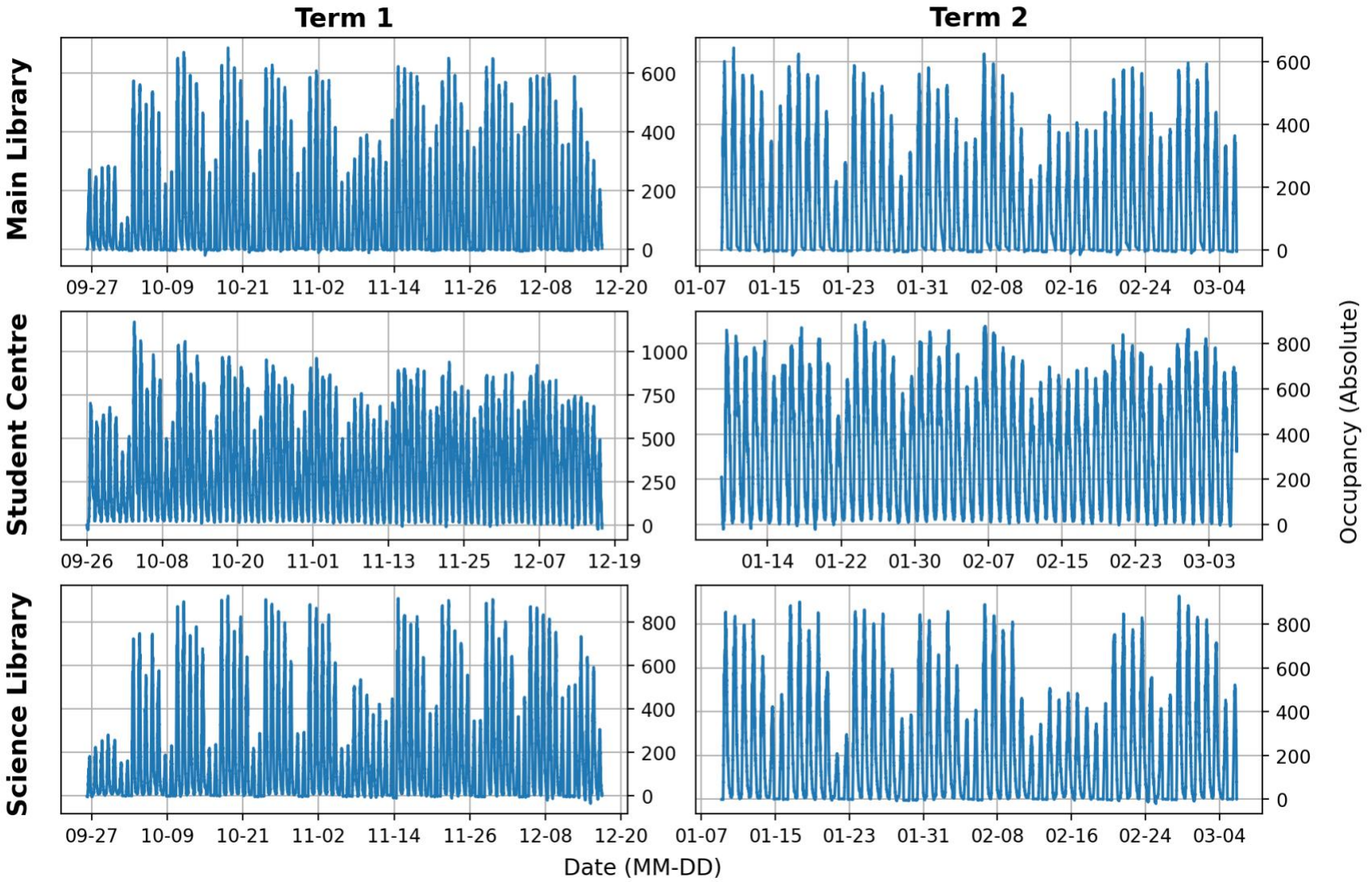

Figure A3: Corrected ML, SC, and SL occupancies for T1 and T2 using manual correction and daily reset

University College London
Torrington Place
LONDON WC1E 7JE

Appendix D

Table A1: Spearman's Rank Correlation Results (P-Values)

Feature	ML		SC		SL	
	T1	T2	T1	T2	T1	T2
Time of Day	0	0	0	0	0	0
Day of Week	0	0	0	0	0	0
Week of Term	0.142	0	0	0.002	0.058	0.125
Reading Week	0.009	0	0	0	0	0
Induction Week	0.011	0	0	0	0	0
Alternative Occupancy 1	SC: 0	SC: 0	ML: 0	ML: 0	ML: 0	ML: 0
Alternative Occupancy 2	SL: 0	SL: 0	SL: 0	SL: 0	SC: 0	SC: 0

Appendix E

Loss Functions in Machine Learning Regression Models

As machine learning regression models can be complex, model characteristics such as the method, features, and loss functions need to be chosen carefully. Even with the same method, differing loss functions can easily vary model performance, defining rewards and penalties differently when evaluating errors within the model.

During training, the model attempts to improve its accuracy by minimising a loss function such as mean squared error (MSE) or mean absolute error (MAE). While both functions are symmetric, treating positive and negative errors equally, MSE is more sensitive to outliers as larger errors are penalised more heavily (6). In contexts where over or under-predictions are viewed as relatively less desirable, a model can employ a weighted loss function. In weighted functions, positive and negative errors are no longer treated equally and are instead characterised by an asymmetrical function.

University College London
Torrington Place
LONDON WC1E 7JE

Appendix F

Occupancy of Alternative study Spaces as a Machine Learning Feature

While occupancies of alternative spaces as a machine learning feature have limited practical applications, models that use these features are far more accurate and consistent. This is due to the large correlation between the area of interest's occupancy and the occupancy of other buildings. Adding these features to the training set improves the performance of all methods, most notably MLR and SVR. However, as RF still performs the best with the lowest RSME and highest R^2 values, this investigation also advises that RF should be pursued as the final machine learning method.

University College London
Torrington Place
LONDON WC1E 7JE

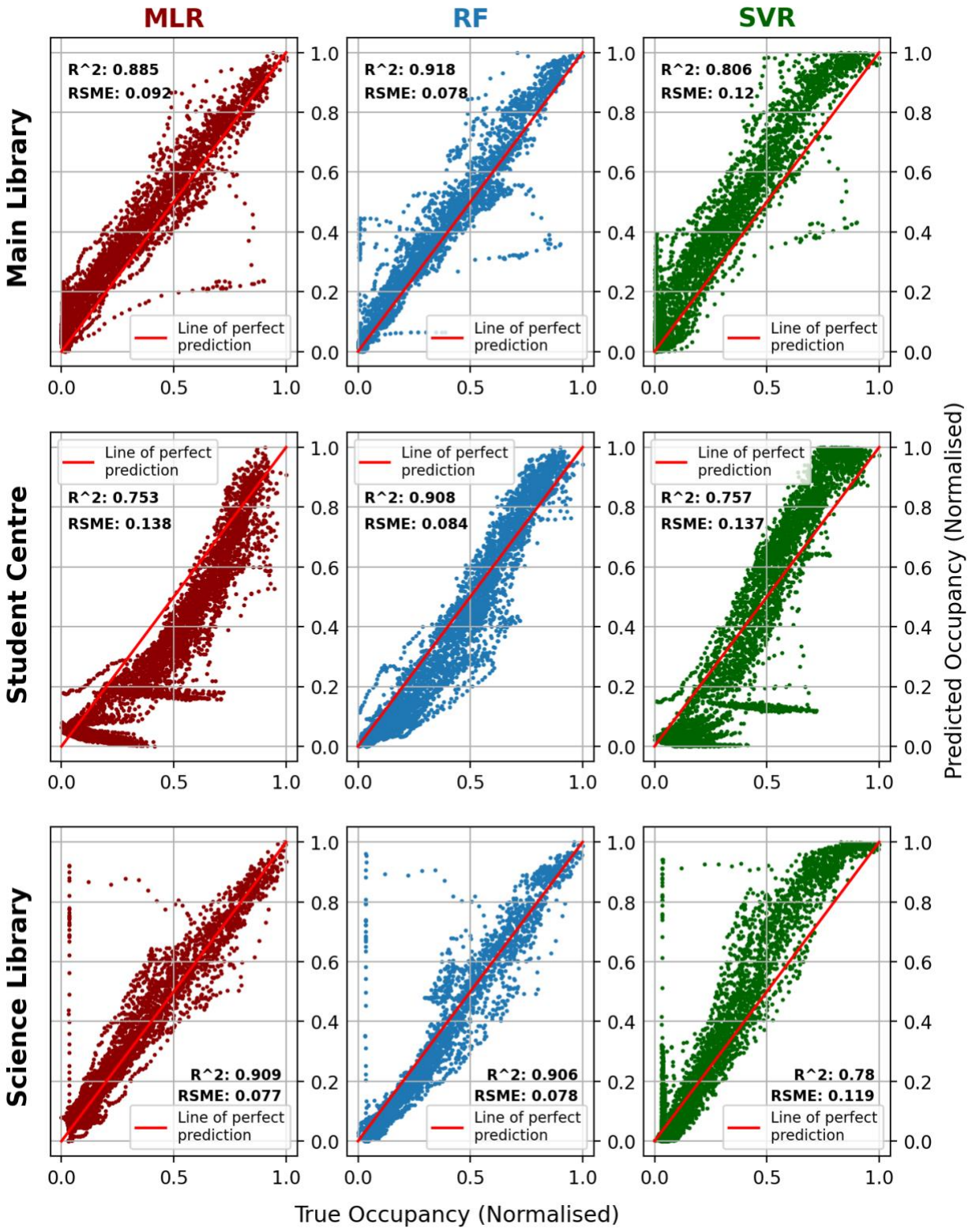

Figure A4: T1 predicted vs. true occupancy (Weeks 1-7 train, 8-12 test)

University College London
Torrington Place
LONDON WC1E 7JE

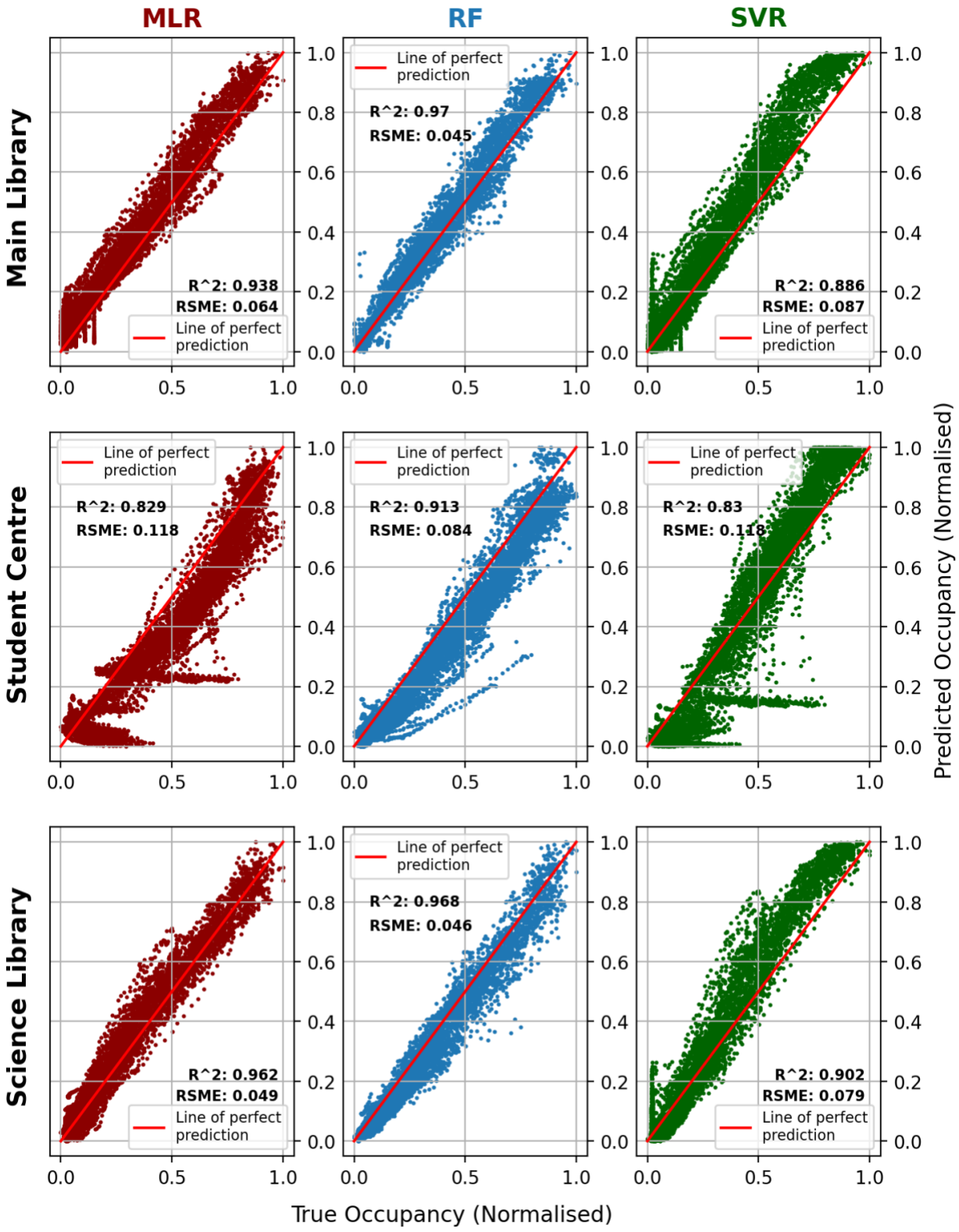

Figure A5: T2 predicted vs. true occupancy (T1 train)

University College London
Torrington Place
LONDON WC1E 7JE

Appendix G

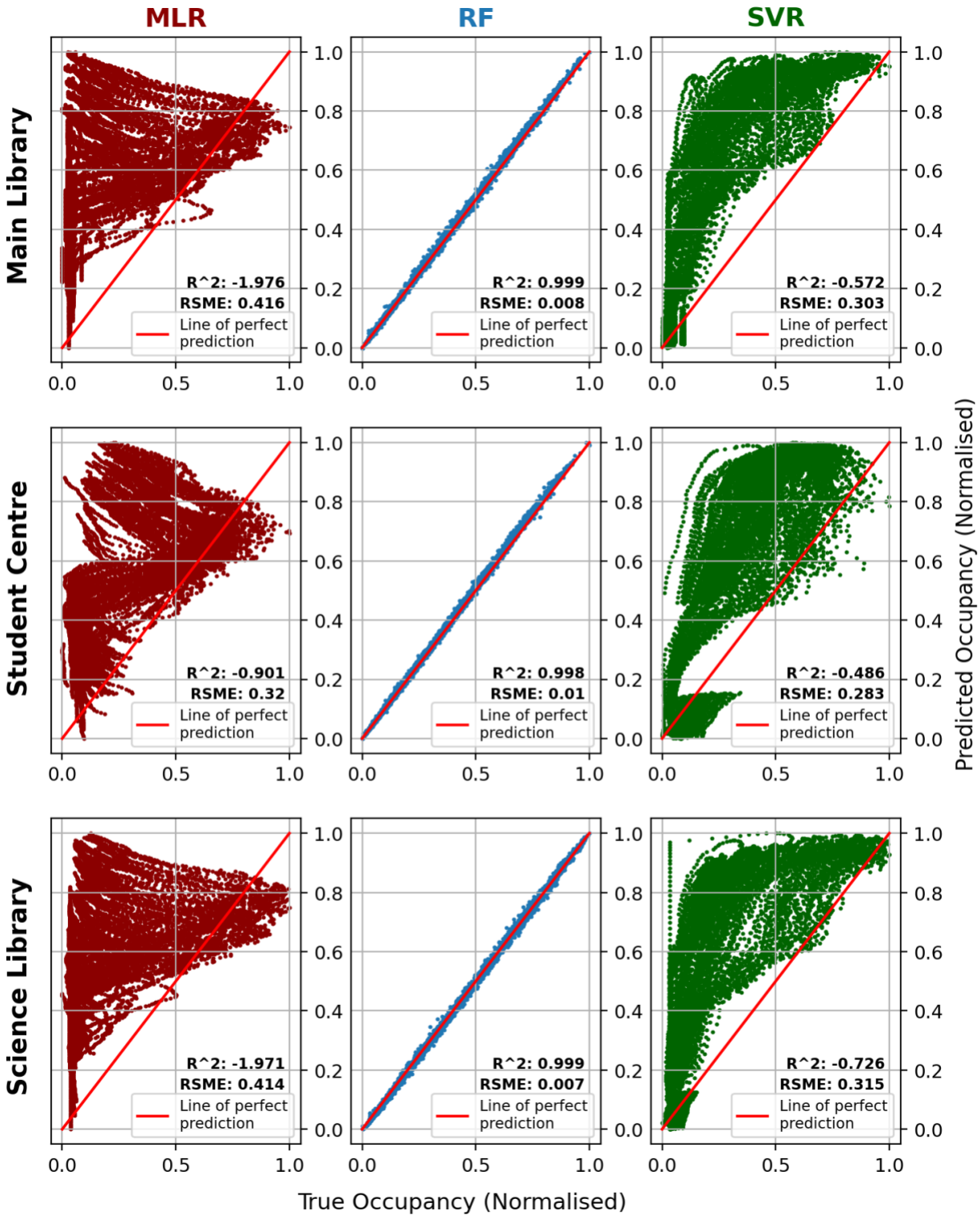

Figure A6: T1 predicted vs. true occupancy (T1 train, self-validation)

University College London
Torrington Place
LONDON WC1E 7JE

Appendix H

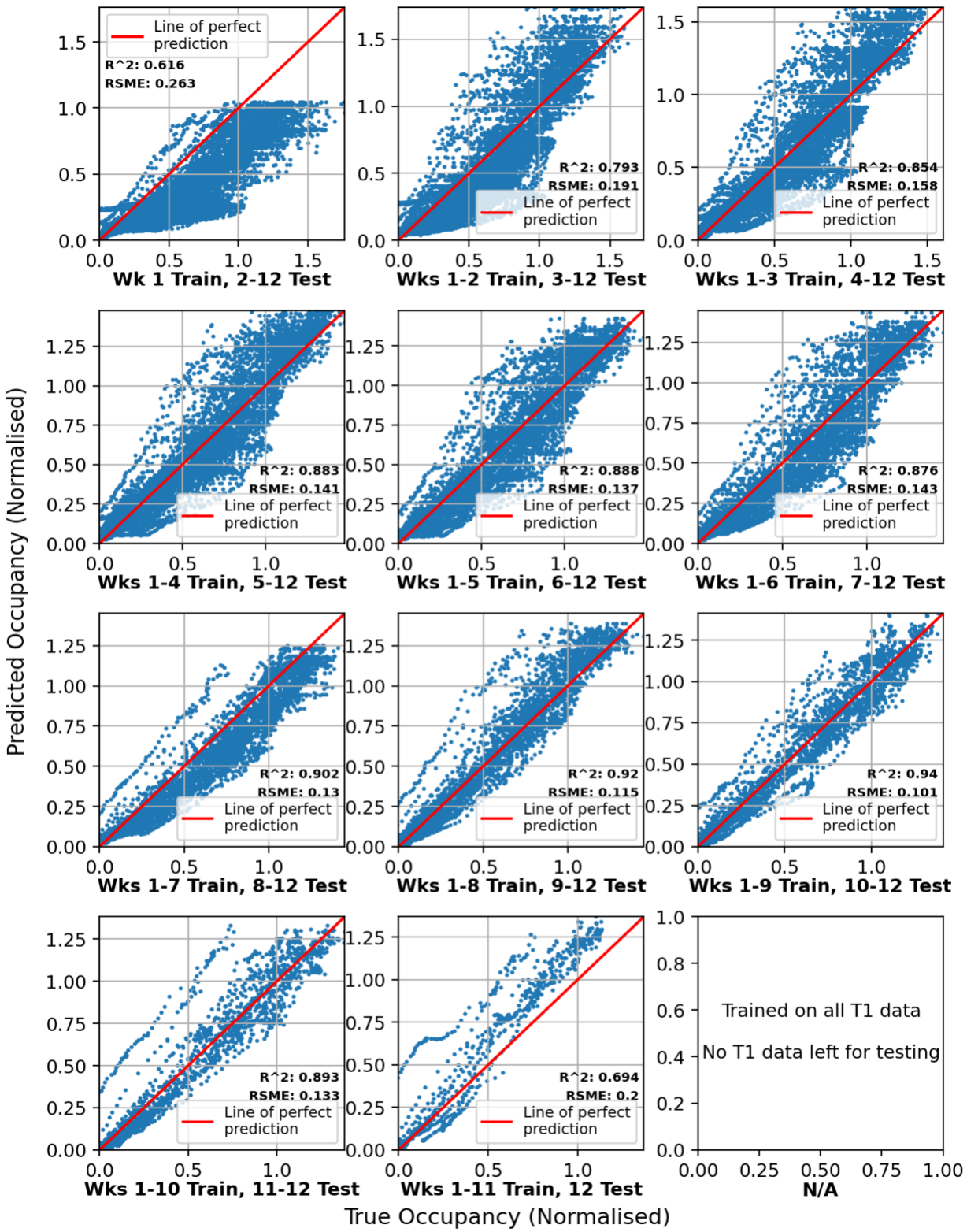

Figure A7: Predicted vs true T1 occupancy using incrementally trained RF models (SC)

University College London
Torrington Place
LONDON WC1E 7JE

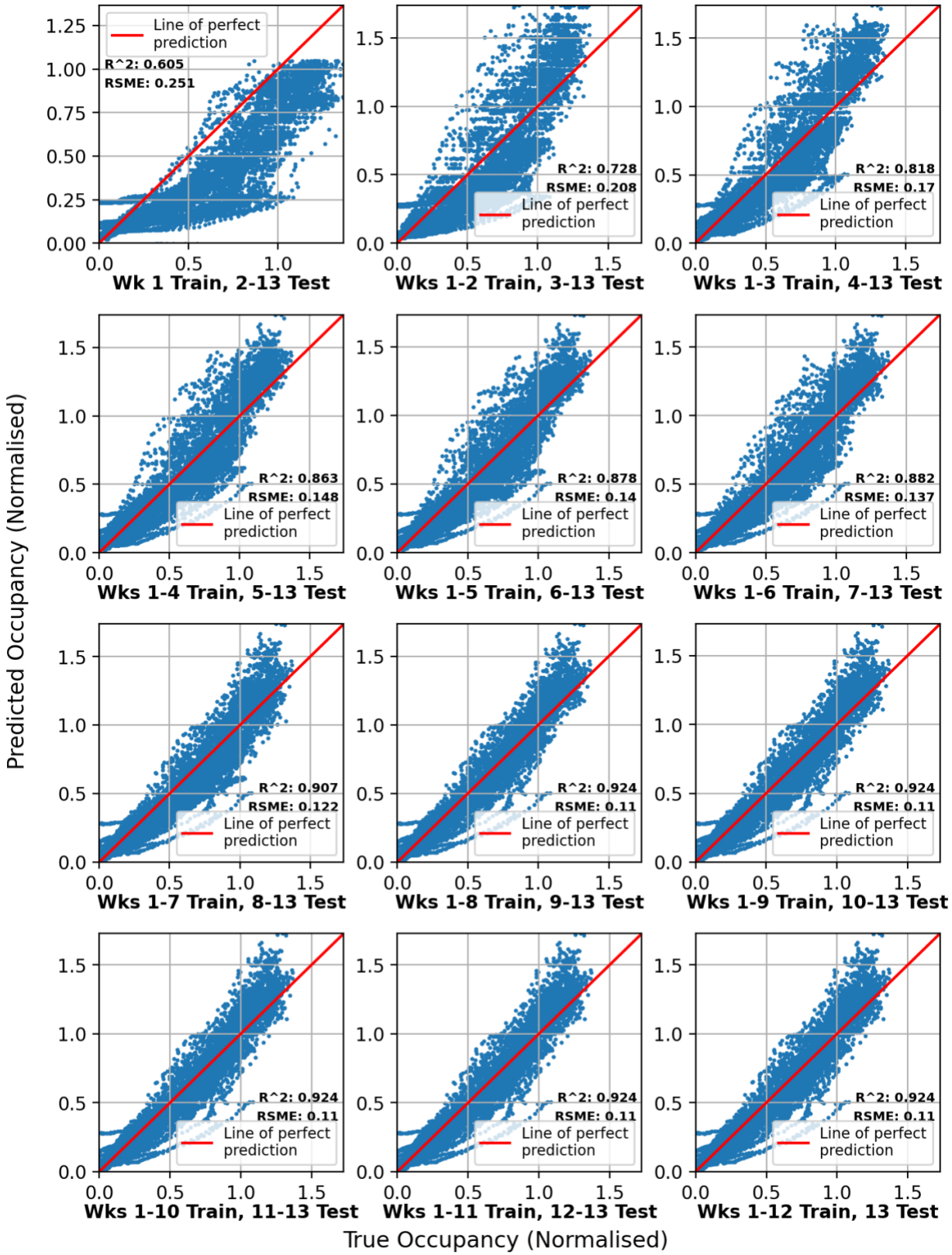

Figure A8: Predicted vs true T2 occupancy using incrementally trained RF models (SC)

University College London
Torrington Place
LONDON WC1E 7JE

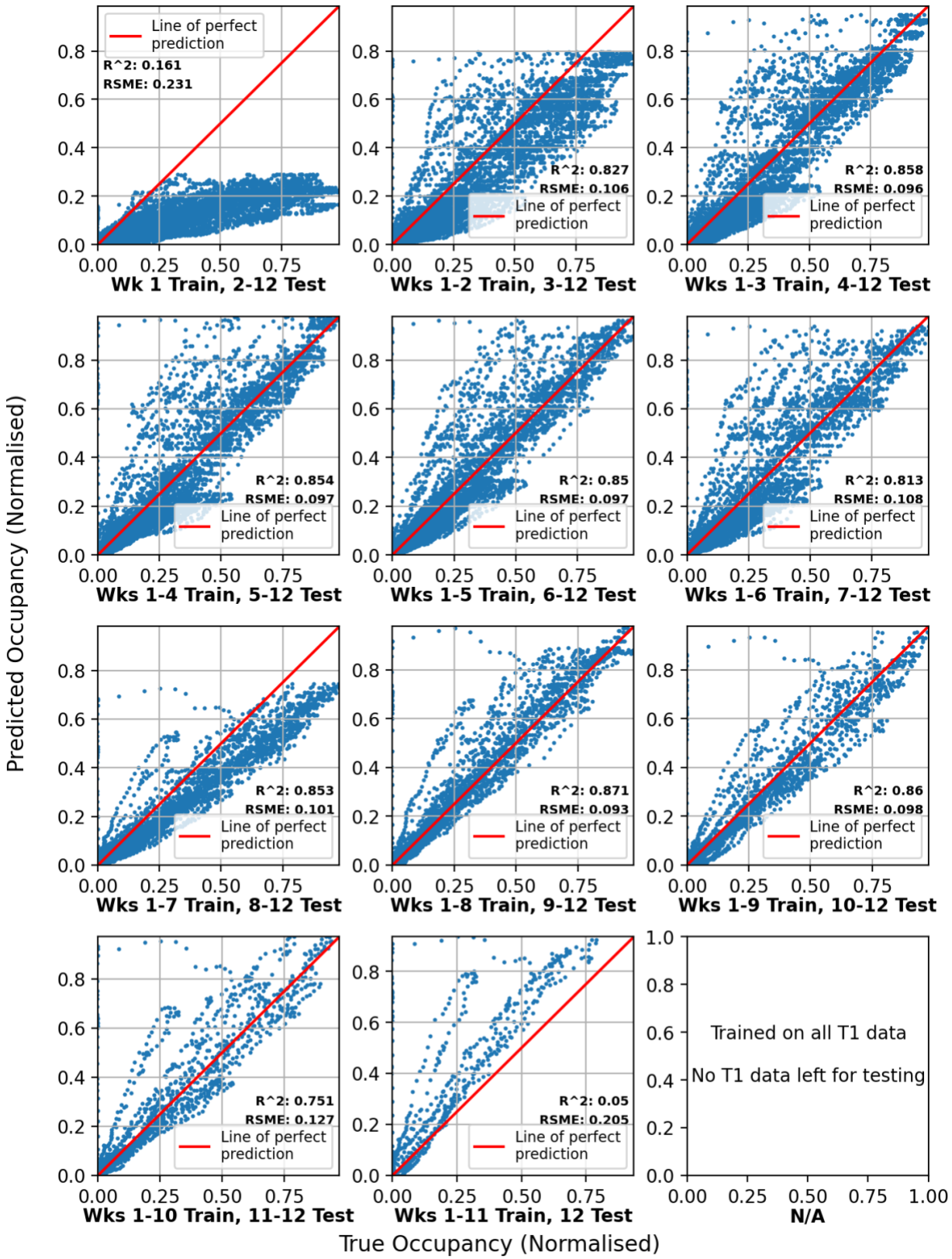

Figure A9: Predicted vs true T1 occupancy using incrementally trained RF models (SL)

University College London
Torrington Place
LONDON WC1E 7JE

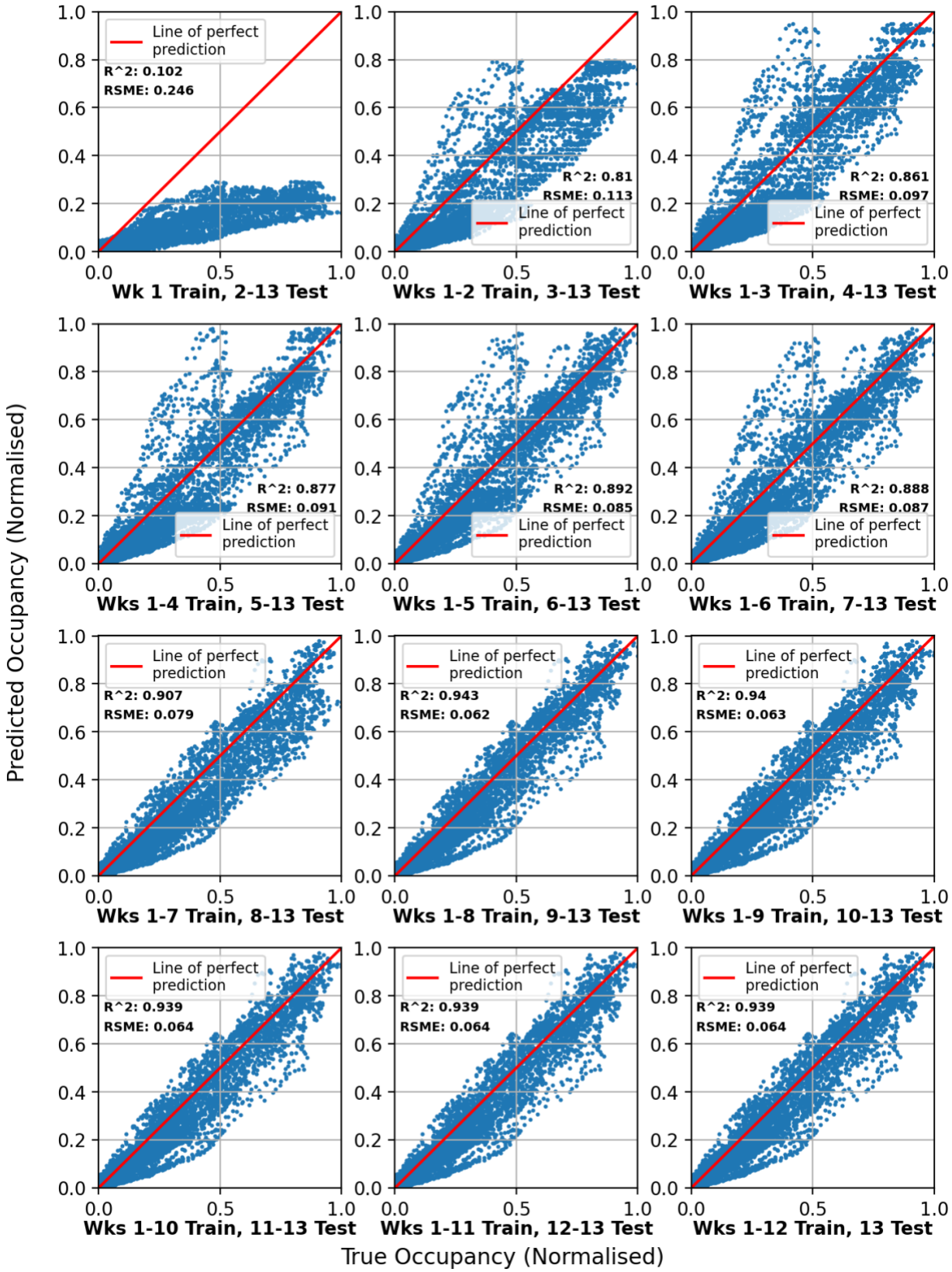

Figure A10: Predicted vs true T2 occupancy using incrementally trained RF models (SL)

University College London
Torrington Place
LONDON WC1E 7JE

Appendix I

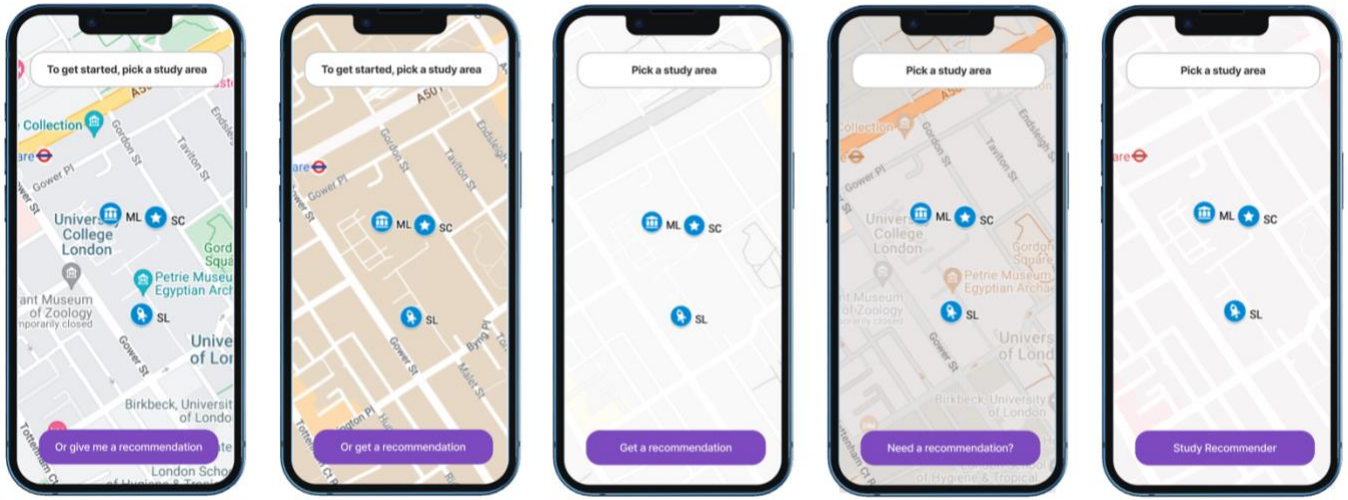

Figure A11: Home-screen concept drafts, varying map background (Google Maps) and word choice

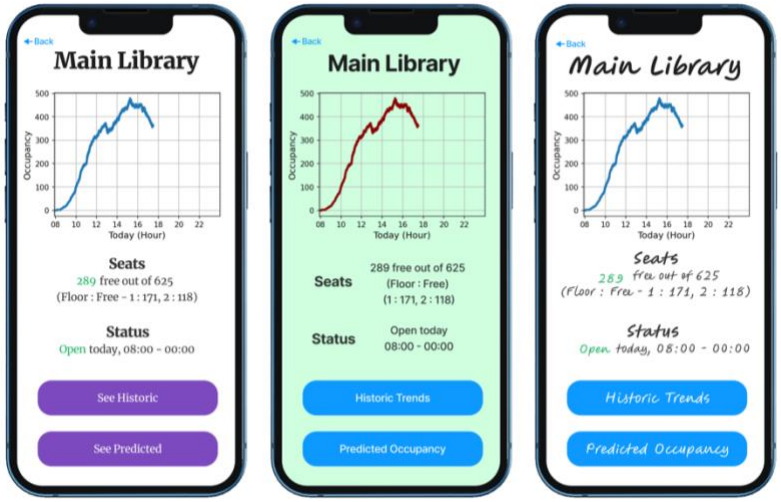

Figure A12: UX/UI concept drafts, varying colour, font, text size, and placement

University College London
Torrington Place
LONDON WC1E 7JE

Appendix J

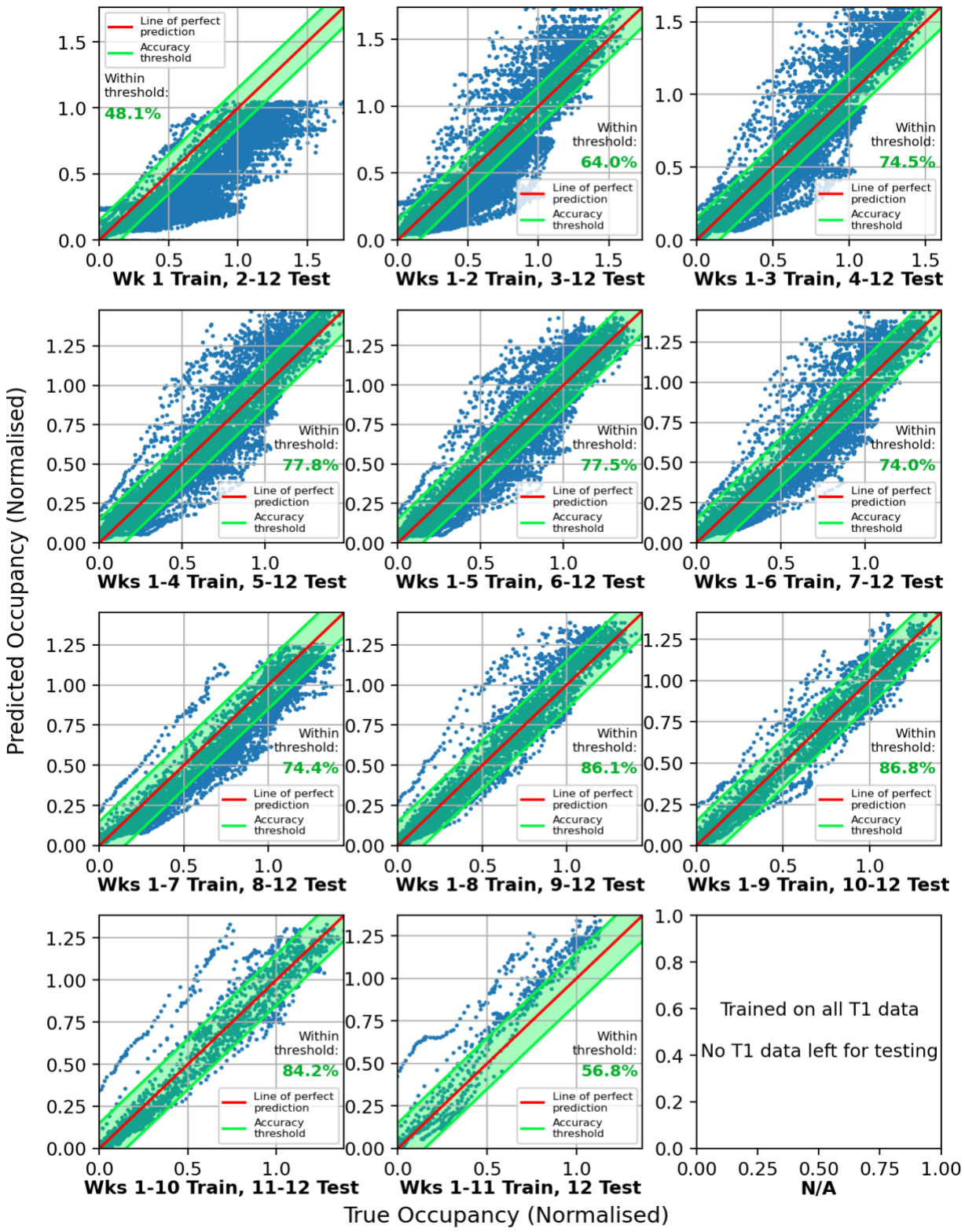

Figure A13: Predicted vs true T1 occupancy using incrementally trained RF models (SC)

University College London
Torrington Place
LONDON WC1E 7JE

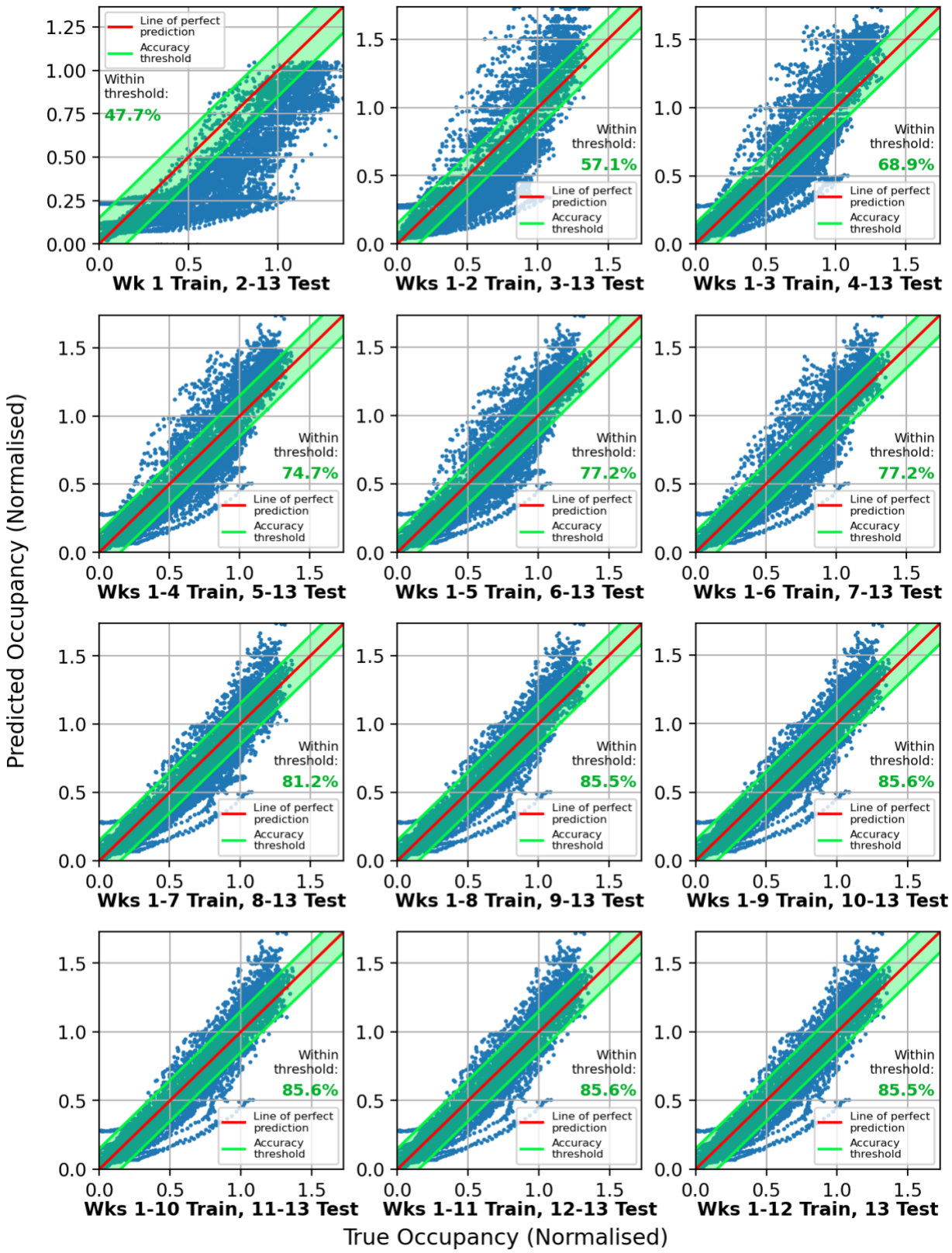

Figure A14: Predicted vs true T2 occupancy using incrementally trained RF models (accuracy threshold) (SC)

University College London
Torrington Place
LONDON WC1E 7JE

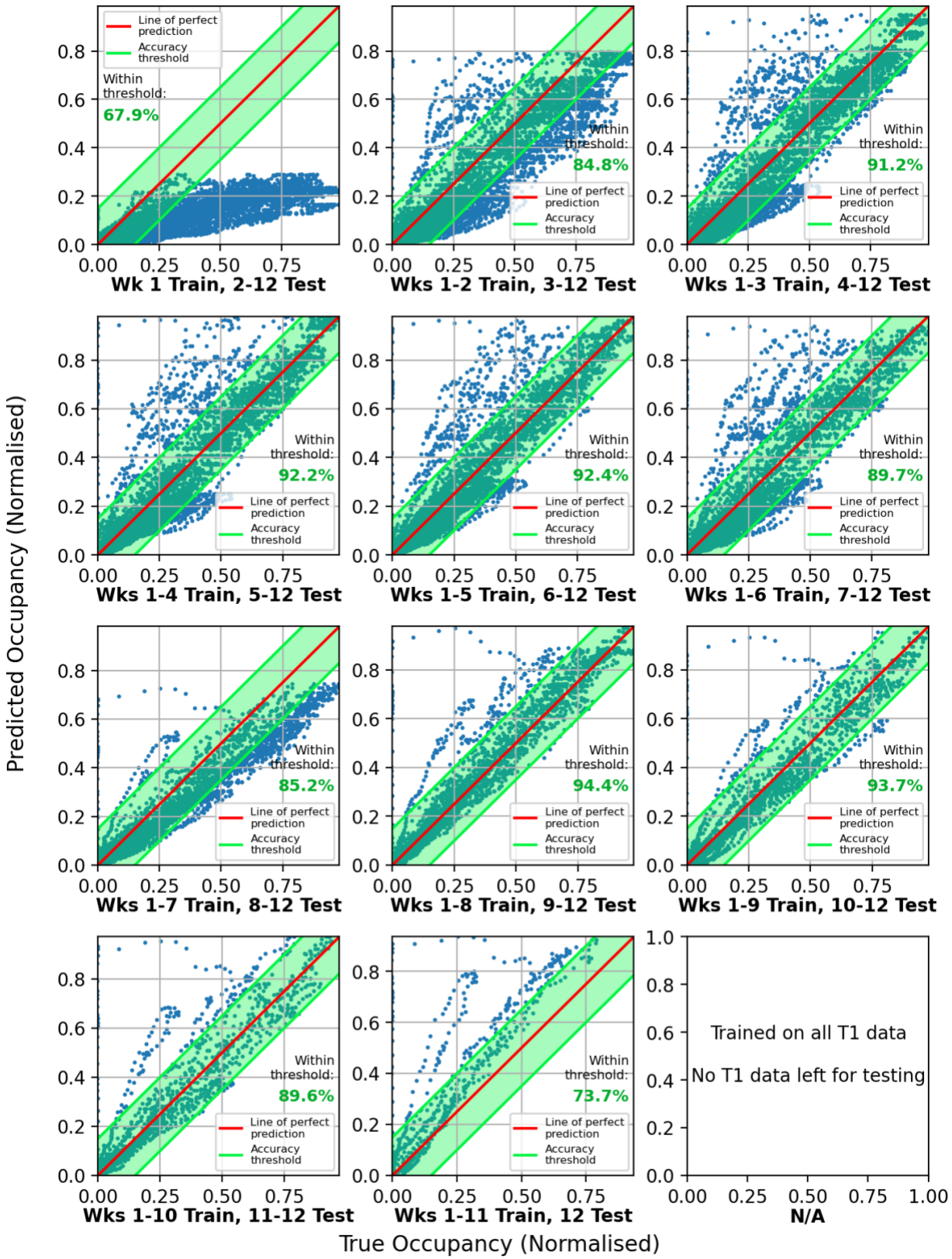

Figure A15: Predicted vs true T1 occupancy using incrementally trained RF models (accuracy threshold) (SL)

University College London
Torrington Place
LONDON WC1E 7JE

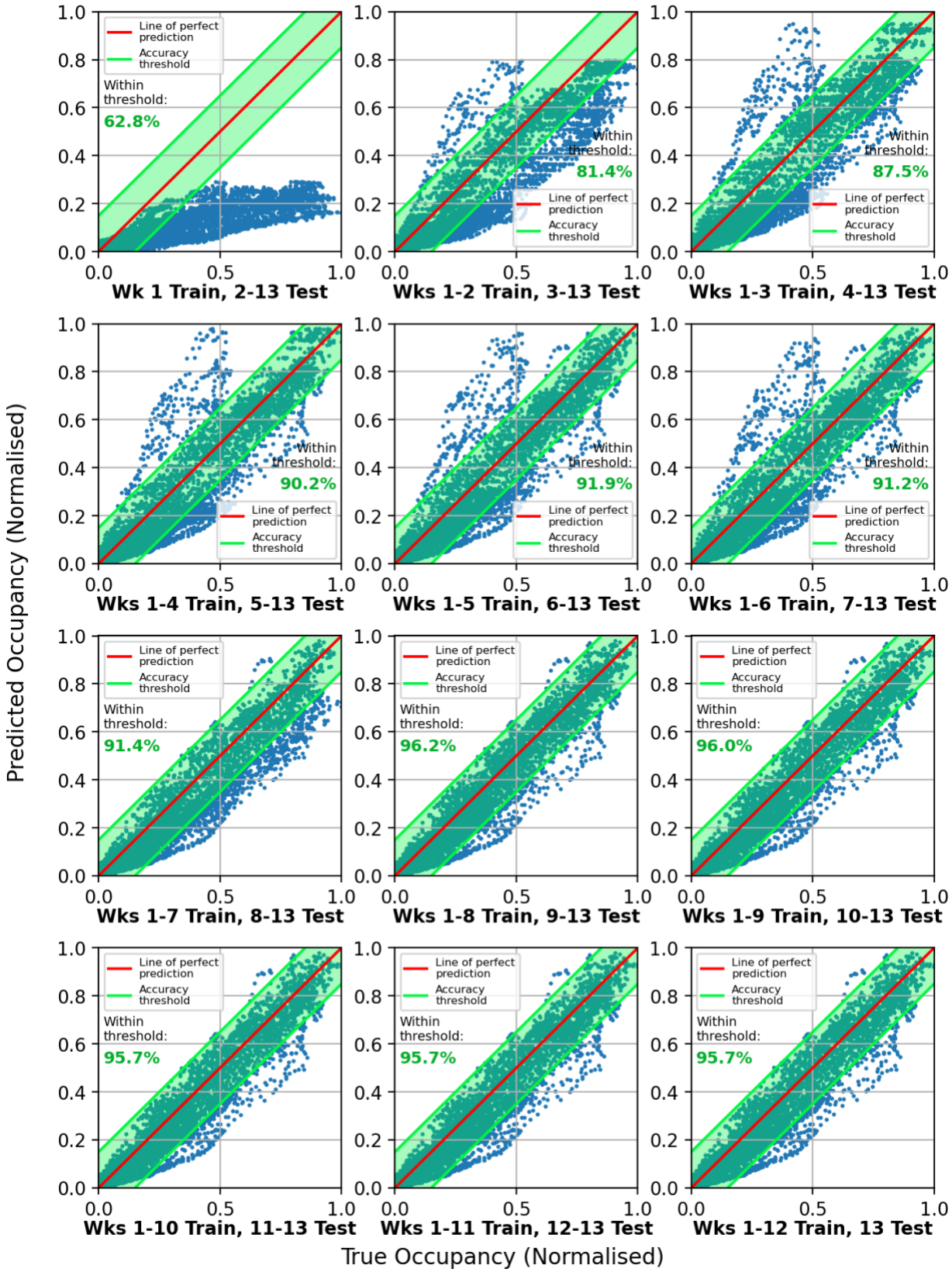

Figure A16: Predicted vs true T2 occupancy using incrementally trained RF models (accuracy threshold) (SL)

University College London
Torrington Place
LONDON WC1E 7JE

Appendix K

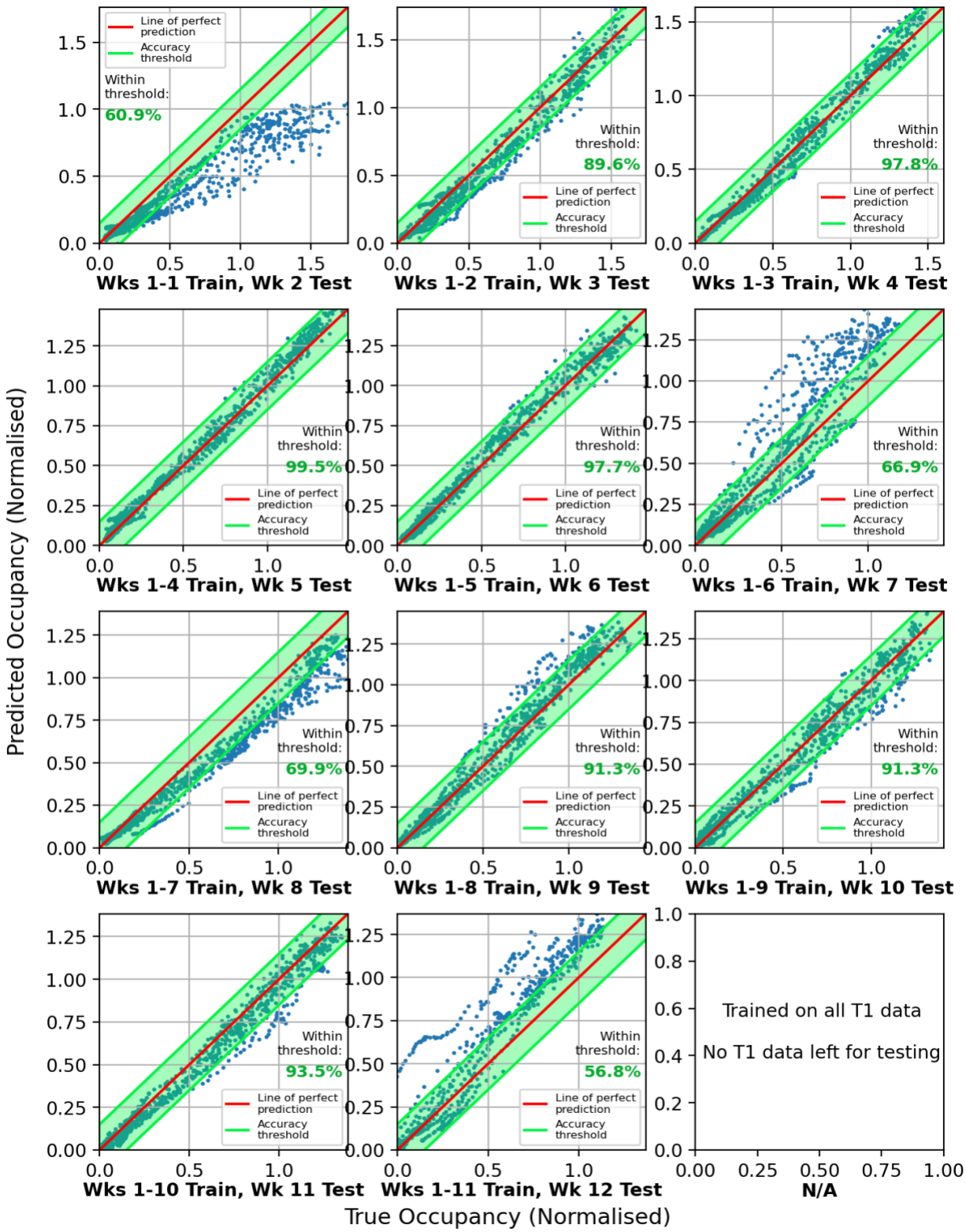

Figure A17: Predicted vs true T1 occupancy using incrementally trained RF models (1-week test sets) (SC)

University College London
Torrington Place
LONDON WC1E 7JE

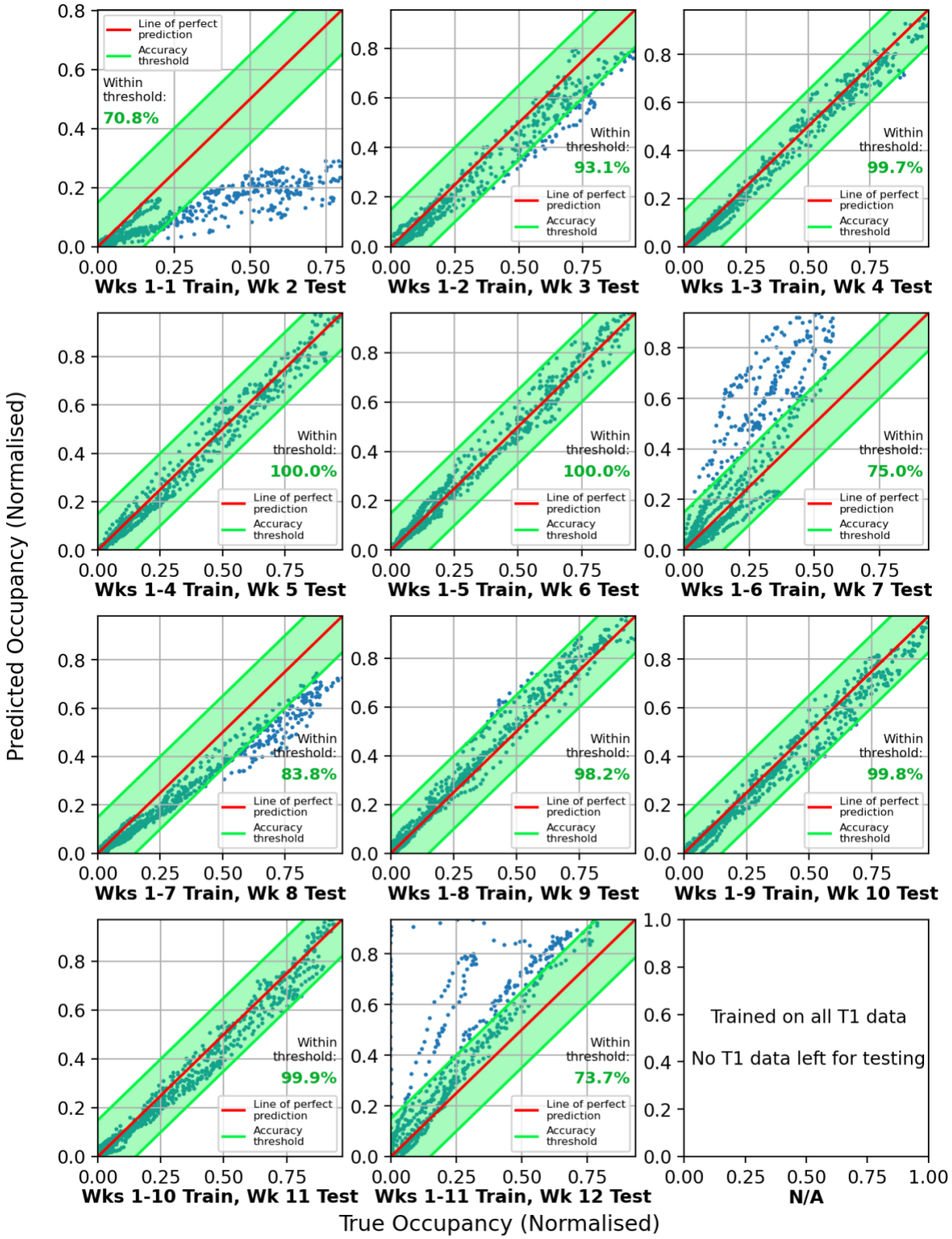

Figure A18: Predicted vs true T2 occupancy using incrementally trained RF models (1-week test sets) (SL)

University College London
Torrington Place
LONDON WC1E 7JE