Cite as: Tekler, Z. D., Ono, E., Peng, Y., Zhan, S., Lasternas, B., & Chong, A. (2022, December). ROBOD, room-level occupancy and building operation dataset. In Building Simulation (Vol. 15, No. 12, pp. 2127-2137). Tsinghua University Press. doi: https://doi.org/10.1007/s12273-022-0925-9

ROBOD, Room-level Occupancy and Building Operation Dataset

Zeynep Duygu Tekler[1], Eikichi Ono[1], Yuzhen Peng[1], Sicheng Zhan[1], Bertrand Lasternas[1], and Adrian Chong[1,*]

[1]Department of the Built Environment, National University of Singapore, 4 Architecture Drive, 117566, Singapore
[*]corresponding author(s): Adrian Chong (adrian.chong@nus.edu.sg)

ABSTRACT

The availability of the building's operation data and occupancy information has been crucial to support the evaluation of existing models and development of new data-driven approaches. This paper describes a comprehensive dataset consisting of indoor environmental conditions, Wi-Fi connected devices, energy consumption of end uses (i.e., HVAC, lighting, plug loads and fans), HVAC operations, and outdoor weather conditions collected through various heterogeneous sensors together with the ground truth occupant presence and count information for five rooms located in a university environment. The five rooms include two different-sized lecture rooms, an office space for administrative staff, an office space for researchers, and a library space accessible to all students. A total of 181 days of data was collected from all five rooms at a sampling resolution of 5 minutes. This dataset can be used for benchmarking and fostering data-driven approaches in the field of occupancy prediction and occupant behaviour modelling, building simulation and control, energy forecasting and various building analytics.

Background & Summary

The building sector is currently responsible for more than one-third of the global energy consumption and approximately 40% of the total direct and indirect CO_2 emissions in the world[1]. As energy demand from the building sector continues to rise due to rapid urbanisation around the globe, significant efforts have been dedicated to improving building energy efficiency while maintaining reliable building operations and high indoor environmental quality for the occupants.

To achieve this goal, researchers have relied on various modelling approaches and simulation tools to help model and quantify building energy use based on different factors, such as climatic regions[2], architectural design[3], environmental conditions[4], occupancy and occupant interactions with building systems[5,6]. These models often require the systematic collection and analysis of various real-world inputs such as the buildings' operational data, energy use and occupancy information to derive meaningful insights and develop effective strategies for reducing building energy use[7]. For instance, the availability of various building data is necessary for physics-based energy models to define model assumptions and inform model calibration[8]. At the same time, data-driven or machine learning-based models require a sufficient amount of training data to produce reliable prediction results[9,10].

However, the collection of such real-world datasets is often challenging in reality. Firstly, it requires the installation of different sensors within each room in the building, which can incur a considerable cost depending on the number of rooms and the size of the target building. After the sensors have been deployed, another significant cost comes from the regular maintenance of these sensors to ensure they stay operational and the data storage services procured to safely store and manage the data collected. Secondly, the integration of the sensor data collected can also create additional hurdles due to the issues related to intermittent sensor failure and nonstandard sampling frequencies used by different sensor manufacturers. Lastly, the collection of occupancy data, which is often performed in person or through surveillance cameras, is labour intensive and may also encounter resistance from the building occupants due to privacy concerns[11]. Despite these challenges, there has been a sustained effort within the building science community to encourage the release of public building datasets to facilitate collaborative and reproducible research. Some examples of these public datasets include: the Building Data Genome Project 2[12], which contains the energy metering data for 1,636 non-residential buildings; BLOND[13], an energy consumption dataset for appliances in an office building; flEECe[14], an energy use and occupant behaviour dataset for residential buildings; CU-BEMS[15], which contains the electrical consumption and indoor environmental sensor data for a smart office building; as well as other commercial and residential datasets containing energy consumption data, building operation data, occupancy data, indoor environmental quality data or different combinations of these data categories[16–18]. So far, no public dataset contains all the highlighted data categories alongside high-resolution occupancy data across multiple room types.

In this paper, we release ROBOD, a Room-level Occupancy and Building Operation Dataset. To the best of our knowledge, this is the most comprehensive dataset that contains room-level occupant presence and count information integrated with building operation data from different room types in a university environment. The dataset consists of a wide range of data categories, including indoor environmental conditions, Wi-Fi connected devices, building energy end-uses (i.e., HVAC, lighting, plug loads, and fans), HVAC operations, and local outdoor weather conditions collected through various heterogeneous sensors together with the ground truth occupant presence and count information for five different rooms. Through the use of this dataset, researchers from different fields can benefit from various applications, including but not limited to occupancy prediction and occupant behaviour modelling, building simulation and control, energy forecasting, and building analytics.

Methods

Building & Room Characteristics

The building considered in this dataset is the School of Design and Environment 4 (SDE4) building located at the National University of Singapore. SDE4 is a 6-story academic building spanning 8,588 square meters. It is the first newly-built net-zero energy building in Singapore and the first building in South Asia that obtained a Zero Energy Certification. We collected the room occupancy and building operation data for five rooms located at different building levels, as visualized in Figure 1. The five rooms include two different-sized lecture rooms (Room 1 and Room 2), an office space for administrative staff (Room 3), an office space for researchers (Room 4), and a library space for students (Room 5). The detailed description of each room is provided in Table 1.

Data Categories Overview and Collection

A building management system (BMS) is currently deployed in the building to help monitor and manage the building's mechanical and electrical systems. As part of BMS, various sensors are installed throughout the study building to collect information about the building's energy consumption, HVAC conditions and outdoor weather conditions. The BACnet Protocol is used to retrieve these sensor measurement data to be stored in the PI Data Archive, which is a feature within the OSISoft PI system. The PI Data Archive serves as an industry-standard data management system for storing time-series data and allows users to perform remote data extraction using various RESTful API services. On top of that, we have installed standalone indoor environmental quality (IEQ) sensors to measure the indoor environmental conditions within each room. Apart from these data categories, we have also tapped into the surveillance cameras and Wi-Fi access points within the study building to obtain room-level occupancy information and the number of Wi-Fi-connected devices, respectively. All data measurements from different sensors were queried with a sampling frequency of 5 minutes before they are integrated to form ROBOD. A 5-minute sampling interval was chosen to strike a good balance between data representativeness and data collection cost.

The following section describes the details of each data category found within the dataset. More detailed information about the data units, sensor types, sensor range and accuracy specifications from the manufacturers are provided in Table 2.

Indoor Environmental Quality Data

The indoor environmental data represent the measurements for indoor environmental quality, which include VOC (volatile organic compound), sound pressure level, relative humidity, indoor air temperature, illuminance, PM2.5 (particulate matter), and CO_2 concentration levels. A dedicated IEQ monitoring unit is installed in each room and its location within the room is provided in Table 3.

Wi-Fi Data

The Wi-Fi data represents the number of Wi-Fi-enabled devices connected to the routers installed in each room. Some examples of these devices include mobile devices (i.e., smartphones and laptops), which connect to the nearest routers depending on their users' movement patterns and location, and stationary devices whose location remains fixed mainly within the room (i.e., printers and desktops). Based on this, number of Wi-Fi connected devices are higher than the number of occupants in the room as the Wi-Fi data contain both mobile and stationary devices. However for further processing, any filtering logic could be introduced to filter out the stationary devices within the room to identify the number of mobile devices which could be used to infer the occupant count. The raw Wi-Fi dataset contains the logs for every device that connects to different access points across the campus and is stored in a Hive SQL database. By querying the relevant logs through the Open Database Connectivity (ODBC) API, the raw Wi-Fi logs are processed to extract the number of connected devices by counting the number of unique MAC addresses recorded during a 5-minute interval for each room.

Energy Data

The energy data represents the energy consumption values of the building's electrical end uses such as HVAC, lighting, plug loads, and ceiling fans. For HVAC energy consumption,

- Room 1 and Room 2 are conditioned by Fan Coil Units (FCU), with the chilled water supplied by a district chiller plant and the supply airflow rate controlled by variable speed fans.

- Room 3, Room 4, and Room 5 are conditioned by Air Handling Units (AHU), which are connected to multiple rooms in the building. It should be noted that the chilled water energy and AHU fan energy consumption are same for the rooms which share the same AHU. In this dataset, Room 4 and Room 5 share the same AHU and therefore have identical energy consumption values for chilled water and AHU fan energy.

The energy consumption data of lighting, plug loads, and ceiling fans are collected through electrical meters and the number of each end use (i.e, lighting, plug loads and ceiling fans) found in each room is listed in Table 4. In this case, the number of lighting units refers to the number of luminaries in each room. Similarly, plug load units are represented as the number of inbuilt 13A double electrical sockets available in each room. It is also worth highlighting that each room may contain different types of plug loads depending on its space function. For instance, Room 1 and Room 2 contain mostly laptops and projectors, Room 3 and Room 4 contain different number of monitors, laptops, desktops, and printers; while Room 5 contains mostly laptops and printers.

HVAC Operations Data

The HVAC operations data represent the different parameters and settings that the building's HVAC system operates within. Some of these measurements include supply airflow, damper position, temperature setpoint, cooling coil valve position and cooling coil valve command, AHU/FCU fan speed, offcoil air temperature, offcoil temperature setpoint, supply air humidity, pressure across filter, supply air static pressure and supply air temperature. It should be noted that the building uses a dedicated outdoor air system for air supply, so the CO_2 level of incoming air is identical to the outdoor CO_2 level.

The temperate setpoint in all rooms is conditioned by Proportional Integral Derivative (PID) control against the thermostat temperature setpoint set by the room occupants. As Room 1 and Room 2 are conditioned by FCUs, they do not contain data measurements related to VAV. The availability of the HVAC operations is also indicated in Table 2 as a footnote.

- Room 1 and Room 2 have dedicated FCUs supplying airflow rates at 3,375 and 2,025 cubic meter per hour (CMH), respectively.

- Room 3 has a VAV airflow rate of 900 CMH and is air-conditioned by a AHU with a supply airflow rate of 1,3470 CMH, serving five other rooms in the building. Room 4 has a VAV airflow rate of 3,192 CMH, while Room 5 has a VAV airflow rate of 1,944 CMH. Both rooms are air-conditioned by the same AHU with a supply airflow rate of 14,560 CMH, supplying chilled air to eleven other rooms in the building.

Outdoor Weather Data

The outdoor weather data is measured by a local weather station installed on the roof of the study building. Measurements include barometric pressure, dry bulb temperature, global horizontal solar radiation, wind direction and speed, outdoor CO_2, and relative humidity.

Occupancy Data

The occupancy data contains both the occupant presence and number of occupants present in each room. This information was collected by monitoring the occupants' movement through surveillance camera footage and manually counting the number of occupants. At any point in time during the data collection process, any identifiers (i.e., names and personal details) that reveal occupants' identity were not collected nor stored in this dataset. The protocols for the data collection has been approved by the host university's Institutional Review Board (NUS-IRB-2021-31).

Data Pre-processing

This section details the data pre-processing steps performed when merging the data categories described above to form ROBOD. These steps involve formatting the timestamp information for each data category to follow the same ISO 8601 date-time format (i.e., YYYY-MM-DD HH:MM +-HH:MM), starting with the year information, followed by the month, day, hour, minute, and time zone offset from UTC. Each data measurement follows a 5-minute sampling interval, starting with the 0^{th} minute, followed by the 5^{th} minute, the 10^{th} minute, and so on till the 55^{th} minute during each hour. After these standardisation steps are performed, the six categories are merged within the same timestep using their timestamp information as the primary key.

Data Records

ROBOD is currently hosted on figshare[19] and consists of five comma-separated value (CSV) files. Each file contains the combined data for each room for all six data categories described in Table 2. Each data measurement also contains the timestamp information corresponding to the time when the data measurement was recorded and followed the date-time format: YYYY-MM-DD HH:MM +08:00. The last component (i.e., +08:00) indicates a UTC offset of +8 hours as the data collection was conducted in the tropical island of Singapore. Given that the data measurements followed a sampling interval of 5 minutes, this corresponds to 288 data points recorded per day. The data collection period spanned between September 2021 and December 2021, where the sensor data collected during the weekends were excluded. Furthermore, there were also specific days during the data collection period when several of the sensors were not working correctly for certain rooms, leading to the data collected during these periods being dropped from the final dataset. In the end, a total of 181 days of data was collected from the five rooms, where Room 1, Room 2 and Room 3 contributed 29 days of data separately while Room 4 and Room 5 contributed 47 days of data each. Apart from the timestamp information that is stored in the string format, the occupancy count and presence information is stored as integers, while the rest of the data fields are represented as floating numbers.

Technical Validation

This section presents the technical validity of our dataset starting with a preliminary analysis of missing data and various visualisations involving occupant count, outdoor environmental condition, room air temperature, room temperature setpoint, and energy consumption based on the raw dataset.

Missing data A preliminary analysis of the dataset highlighted a small number of missing data points for each room in ROBOD due to issues related to intermittent sensor failure. Table 5 presents a detailed breakdown of the amount of missing data found in each column and for each room. The temporal relationship of the missing data is also presented in Figure 2. It should be reiterated that the datasets for Room 1 and 2 do not contain columns related to VAV (i.e., Supply Air Flow, Damper Position, Cooling Coil Valve Position and Command, Offcoil Temperature Setpoint, Offcoil Air Temperature, Pressure across Filter, and Supply Air Humidity) as they are conditioned by FCUs, therefore they are not included in the dataset.

Occupant count Figure 3 presents the average occupant count for each room on an average weekday. Based on the occupancy fluctuations, it can be observed that the occupant count patterns differ slightly among different rooms. More specifically, the occupant count for Room 1 and Room 2 experience heavy fluctuations throughout the day compared to other rooms. In particular, we observed three distinct peaks in Room 2 that occur at 11 am, 1 pm, and 3 pm, which can be explained by the block lectures that are regularly scheduled during these periods. Room 3 presents a regular office schedule with the office workers arriving at work between 8 to 10 am and leaving the office at the end of the workday between 6 to 8 pm. The occupants in Room 4 are observed to follow a flexible work schedule where the last departure times for some occupants can stretch late into the night after midnight. Lastly, we can observe a sharp increase in occupancy levels in Room 5 from zero at 9 am, followed by a sharp drop back to zero at 9 pm every day, corresponding with the operational hours of the library space.

Outdoor environmental condition Figure 4 shows the monitored outdoor conditions of dry-bulb temperature, global horizontal solar radiation, relative humidity, and CO_2. As the data was collected from the study building located in the tropic, the outdoor dry-bulb temperature ranges from 22.6°C to 35.5°C, where temperatures tend to rise to higher levels in the afternoon (i.e., 12 pm to 4 pm). The global horizontal solar radiation can reach over 1000 W/m^2 between 11 pm and 3 pm. At the same time, the relative humidity ranges from 40% to 100%, of which over 98% accounts for the primary ratio (25%). The cooling systems process dry-bulb temperature and relative humidity to deliver the required supply air flow to cool down the internal thermal zones within the building, while removing thermal energy generated by the solar radiation. The outdoor CO_2 levels span between 439 ppm to 510 ppm, which is used as the basis of maintaining the indoor CO_2 levels at a standard or comfortable range.

Room temperature setpoint Figure 5 depicts the distributions of temperature setpoints for each of the five rooms. As the occupants' thermal sensation is subjective, the temperature setpoints may differ among the rooms and during different periods of the day. Room 1 and Room 2 show a wide range of temperature setpoints, ranging from 22°C to 27.2°C, and from 22°C to 27.7°C, respectively. Room 4 shifted the setpoints to the range of 25.3°C to 28°C. Unlike the other rooms, Room 3 and Room 5 kept the temperature setpoints consistently at 25°C and 26°C, respectively.

Room air temperature Figure 6 shows a heatmap of the average indoor air temperature or thermal distribution at different time periods during the day for each room. The vertical axis indicates each of the five rooms, and the horizontal axis shows the time of the day. For example, it can be observed that the air temperatures in Room 1, Room 2, and Room 3 tends to be cooler than Room 4 and Room 5. Moreover, air temperatures in the afternoon are also warmer than in the morning for all five rooms.

Energy consumption Figure 7 summarizes energy consumption of space cooling, plug load, and lighting in each room. The cooling energy consumption that combines the energy consumed by chilled water and AHU/FCU fans. Since Room 4 and Room 5 are air-conditioned by the same AHU, their cooling energy consumption is not separated for this analysis. The difference in the cooling demand among the rooms can be explained by the differences in room functions and room area. For instance, Room 1 and Room 2 are used as lecture spaces with similar indoor areas resulting in identical cooling energy consumption and schedules. Similarly, the cooling energy and schedules for Room 3 and Rooms 4+5 are similar in terms of its pattern as all three rooms function as multi-occupant offices (i.e., Room 3 and Room 4). Devices that are connected to electrical sockets can be classified into two groups: non-mobile devices located in the rooms and portable devices. The former contributes 24-hour plug load consumption, including the small energy consumption when the devices enter into idle modes.

The latter only need the electricity from electrical sockets when their owners occupy the rooms. For instance, the plug load consumption in five rooms are nearly constant before 7 am. Furthermore, this consumption in Room 1, Room 2, and Room 5 increased simultaneously from 9 am. Similar to the plug load consumption, the energy consumption of lighting is closely related to occupants' room usages. Therefore, the lighting demand increases from 9 am in most of the rooms.

Usage Notes

The dataset provided in this paper is in the CSV format for all rooms and has a total file size of 20 MB. The CSV data format allows the files to be easily imported by most spreadsheet programs and databases. It is also easy to work with due to its human-readable format and can be readily processed and analysed by most popular programming languages such as Python, Java, Javascript, and R.

Due to the presence of missing data in the dataset, we have also included several data post-processing steps as a reference for researchers who would like to use the existing dataset. These steps involves imputing the dataset's missing or erroneous sensor data by using the *missingpy* imputation library. While different imputation algorithms have been utilised in past studies[20], a Random Forest-based imputation algorithm (i.e., MissForest[21]) is adopted in this case by performing column-wise imputation in an iterative fashion. The algorithm begins by imputing the column with the least number of missing values (i.e., candidate column) and filling the missing values in the remaining columns with an initial guess, such as the column's mean. Following this, a Random Forest (RF) model is trained by setting the candidate column as the output variable and the remaining columns as the model's input for those rows that do not contain missing values in the candidate column. After the RF model has been trained, it is used to impute the missing values in the candidate column before moving on to the next candidate column with the second smallest number of missing values. This process is repeated for each column containing missing values over multiple iterations until the difference between the dataset imputed in the previous round and the newly imputed dataset increases for the first time.

Code availability

All data post-processing steps and visualisations performed in this manuscript are implemented using Python 3.6 and public libraries including Numpy and Pandas for data manipulation, Matplotlib, Seaborn, and Missingno for data visualisation, and Missingpy for data imputation. A step-by-step guide has been compiled within a single Juypter notebook and made available on Github (`https://github.com/ideas-lab-nus/robod.git`).

References

1. Buildings, I. T. Iea: Paris. https://www.iea.org/reports/tracking-buildings-2020, Accessed:2022-01-21 (2020).
2. Orouji, P. *et al.* Atlas of heating: Identifying regional climate-dependent heat demands in residential buildings of iran. In *Building Simulation*, vol. 14, 857–869 (Springer, 2021).
3. Ataman, C. & Dino, İ. G. Performative design processes in architectural practices in turkey: architects' perception. *Archit. Eng. Des. Manag.* 1–15, https://doi.org/10.1080/17452007.2021.1995315 (2021).
4. Aydin, E. E. & Jakubiec, J. A. Sensitivity analysis of sustainable urban design parameters : Thermal comfort , urban heat island , energy , daylight , and ventilation in singapore (2018).
5. Tekler, Z., Low, R. & Blessing, L. Using smart technologies to identify occupancy and plug-in appliance interaction patterns in an office environment. In *IOP Conference Series: Materials Science and Engineering*, vol. 609, 062010, https://doi.org/10.1088/1757-899X/609/6/062010 (IOP Publishing, 2019).
6. Peng, Y., Nagy, Z. & Schlüter, A. Temperature-preference learning with neural networks for occupant-centric building indoor climate controls. *Build. Environ.* **154**, 296–308, 10.1016/j.buildenv.2019.01.036 (2019).
7. Ding, Y. *et al.* Review on occupancy detection and prediction in building simulation. In *Building Simulation*, 1–24 (Springer, 2021).
8. Chong, A., Gu, Y. & Jia, H. Calibrating building energy simulation models: A review of the basics to guide future work. *Energy Build.* **253**, 111533, https://doi.org/10.1016/j.enbuild.2021.111533 (2021).
9. Zhan, S. & Chong, A. Data requirements and performance evaluation of model predictive control in buildings: A modeling perspective. *Renew. Sustain. Energy Rev.* 110835, https://doi.org/10.1016/j.rser.2021.110835 (2021).
10. Peng, Y., Rysanek, A., Nagy, Z. & Schlüter, A. Using machine learning techniques for occupancy-prediction-based cooling control in office buildings. *Appl. Energy* **211**, 1343–1358, 10.1016/j.apenergy.2017.12.002 (2018).

11. Tekler, Z. D., Low, R., Gunay, B., Andersen, R. K. & Blessing, L. A scalable bluetooth low energy approach to identify occupancy patterns and profiles in office spaces. *Build. Environ.* **171**, 106681, https://doi.org/10.1016/j.buildenv.2020.106681 (2020).

12. Miller, C. *et al.* the building data genome project 2, energy meter data from the ashrae great energy predictor iii competition. *Sci. data* **7**, 1–13, https://doi.org/10.1038/s41597-020-00712-x (2020).

13. Kriechbaumer, T. & Jacobsen, H.-A. Blond, a building-level office environment dataset of typical electrical appliances. *Sci. data* **5**, 1–14, https://doi.org/10.1038/sdata.2018.48 (2018).

14. Paige, F., Agee, P. & Jazizadeh, F. fleece, an energy use and occupant behavior dataset for net-zero energy affordable senior residential buildings. *Sci. data* **6**, 1–9, https://doi.org/10.1038/s41597-019-0275-3 (2019).

15. Pipattanasomporn, M. *et al.* Cu-bems, smart building electricity consumption and indoor environmental sensor datasets. *Sci. Data* **7**, 1–14, https://doi.org/10.1038/s41597-020-00582-3 (2020).

16. Tekler, Z. D. *et al.* Near-real-time plug load identification using low-frequency power data in office spaces: Experiments and applications. *Appl. Energy* **275**, 115391, https://doi.org/10.1016/j.apenergy.2020.115391 (2020).

17. Schwee, J. H. *et al.* Room-level occupant counts and environmental quality from heterogeneous sensing modalities in a smart building. *Sci. data* **6**, 1–11, https://doi.org/10.1038/s41597-019-0274-4 (2019).

18. Li, H., Wang, Z. & Hong, T. A synthetic building operation dataset. *Sci. data* **8**, 1–13, https://doi.org/10.1038/s41597-021-00989-6 (2021).

19. Tekler, Z. D. *et al.* Robod, room-level occupancy and building operation dataset. *figshare* https://doi.org/10.6084/m9.figshare.19234530.v5 (2022).

20. Low, R., Tekler, Z. D. & Cheah, L. Predicting commercial vehicle parking duration using generative adversarial multiple imputation networks. *Transp. Res. Rec.* **2674**, 820–831, https://doi.org/10.1177/0361198120932166 (2020).

21. Stekhoven, D. J. & Bühlmann, P. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112–118, https://doi.org/10.1093/bioinformatics/btr597 (2012).

Acknowledgements

This research project is supported by the National Research Foundation, Singapore, and Ministry of National Development, Singapore under its Cities of Tomorrow R&D Programme (CoT Award COT-V4-2020-5). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Ministry of National Development, Singapore.

Author contributions

Zeynep Duygu Tekler: Conceptualisation, Methodology, Software, Validation, Data Curation, Writing - Original Draft. **Eikichi Ono:** Investigation, Writing - Review & Editing. **Yuzhen Peng:** Visualisation, Writing - Original Draft. **Sicheng Zhan:** Visualisation, Writing - Review & Editing. **Bertrand Lasternas:** Resources. **Adrian Chong:** Writing - Review & Editing, Supervision, Funding Acquisition.

Competing interests

The authors declare no competing interests.

Figures & Tables

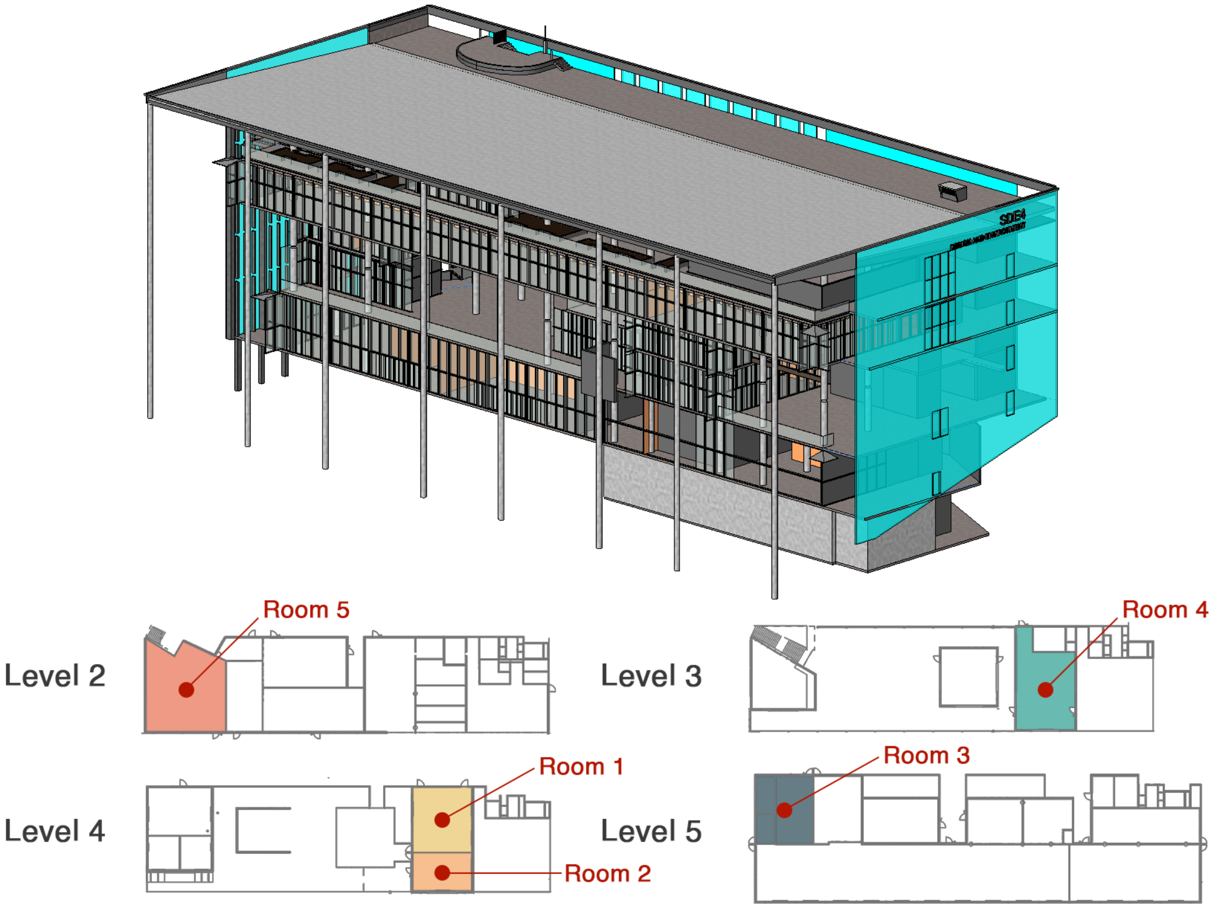

Figure 1. Study building (top) and room layouts corresponding to the building levels (bottom)

Table 1. Room descriptions.

Room	Space Function	Occupant Type	Level	Floor Area [m^2]	Floor to Ceiling Height [m]	Room Volume [m^3]	Seating Capacity [person]	Maximum Occupancy Density [m^2/person]
Room 1	Lecture room	Students	4	118.6	4.1	486.2	40	3.0
Room 2	Lecture room	Students	4	53.7	4.1	220.2	40	1.3
Room 3	Office space	Administrative staff	5	98.4	4.2	413.2	15	6.6
Room 4	Office space	Researchers	3	141.9	4.1	581.7	25	5.6
Room 5	Library space	Students	2	182.8	7.5	1363.3	36	5.0

Table 2. Data categories and sensor specifications.

Data Category	Measured Variable	Data Unit	Sensor Type (Brand)	Sensor Range	Sensor Accuracy
Indoor environmental quality	VOC	ppb	IAQ monitoring unit (Awair Omni)	0 – 60000	±10%
	Sound pressure level	dB(A)		Not specified	Not specified
	Relative humidity	%RH		0 – 100	±2%RH
	Air Temperature	°C		-40 – 125	±0.2 °C
	Illuminance	lux		0 – 64000	Not specified
	PM2.5	$\mu g/m^3$		0 – 1000	$\pm 15 \mu g/m^3$ or ±15%
	CO_2	ppm		400 – 5000	±75 ppm or ±10%
Wi-Fi	Wi-Fi connected devices	Number	Wi-Fi Router (Cisco)	NA[a]	NA[a]
Energy	Ceiling fan energy	kWh	Energy meter (Schneider Electric Acti9 iEM3000)	0 – 999999	±1%
	Lighting energy				
	Plug load energy				
	Chilled water energy		BTU meter (Integra Metering CALEC ST II)		±2%
	AHU/FCU fan energy		Energy meter (Schneider Electric PM5300)		±0.5%
HVAC operations	Supply air flow[b]	CMH	VAV box (Johnson Controls)	0 – 3375	±15%
	Damper position[b]	%		0 – 100	NA[a]
	Temperature setpoint	°C	NA[a]	NA[a]	NA[a]
	Cooling coil valve position[b]	%	Valve (Johnson Controls)	0 – 100	NA[a]
	Cooling coil valve command [b]				
	AHU/FCU fan speed	Hz	Variable speed drive (ABB)	0 – 50	±0.2%
	Offcoil air temperature[b]	°C	NTC thermistor (GreyStone TSDC series)	-40 – 60	±0.2 °C
	Offcoil temperature setpoint[b]	°C	NA[a]	NA[a]	NA[a]
	Supply air humidity[b]	%RH	Capacitive (GreyStone HSDT series)	0 – 100	±2%RH
	Pressure across filter [b]	Pa	Capacitive (Setra Model 264)	Not specified	±1%
	Supply air static pressure				
	Supply air temperature	°C	NTC thermistor (GreyStone TSAP series)	-40 – 60	±0.2 °C
Outdoor weather[c]	Baromatic pressure	hPa	Piezoresistive	600 – 1100	±0.5 hPa @20 °C
	Dry bulb temperature	°C	Pt100	-40 – 60	±0.15°C ±0.1%
	Global horizontal solar radiation	W/m^2	Thermopile	0 – 2000	2nd class pyranometer
	Wind direction	° (Degree)	Ultrasonic	0 – 360	±2° RMSE
	Wind speed	m/s	Ultrasonic	0 – 60	±0.2 m/s or ±2%
	CO_2	ppm	Non-dispersive infrared	0 – 2000	±(5 ppm + 2%)
	Relative humidity	%RH	Capacitive	0 – 100	±1.5%RH
Occupancy	Occupant presence	Binary (1/0)	Surveillance camera (XeronVision 2M HD IP Vari-Focal Lens Dome)	NA[a]	NA[a]
	Occupant count	Number			

[a] NA refers to "Not Applicable".
[b] Indicated measurements are not applicable for Room 1 & Room 2.
[c] All the outdoor weather data were collected by a weather station (Delta OHM HD52.3D).

Table 3. Locations of retrofitted sensors (i.e., surveillance camera and IEQ units).

Room	Surveillance Camera	IEQ Units
Room 1	Two surveillance cameras outside of two doors	Mounted to an east side column vertically on wall
Room 2	One surveillance camera outside of the door	Mounted to an east side column vertically on wall
Room 3	On the top corner of the rooms near the entrance doors	Mounted to a west side column vertically on wall
Room 4	On the top corner of the rooms near the entrance doors	Mounted to a west side column vertically on wall
Room 5	Three surveillance cameras inside the room	Mounted to a east side column vertically on wall

Table 4. Number of end uses in each room.

Room	No. of Ceiling Fans	No. of Luminaries	No. of 13A Double Sockets
Room 1	6	20	26
Room 2	4	12	14
Room 3	4	14	9
Room 4	6	32	20
Room 5	6	11	12

Table 5. A detailed breakdown of the amount of missing data in the relevant columns of each room.

Room	Total	Missing Data	Column Name
Room 1	8352	9	supply_air_pressure and ahu_fan_speed
		10	chilled_water_energy and ahu_fan_energy
		14	voc, sound_pressure_level, indoor_relative_humidity, illuminance, pm2.5, indoor_co2
Room 2	8352	30	voc, sound_pressure_level, indoor_relative_humidity, illuminance, pm2.5, indoor_co2
Room 3	8352	13	voc, sound_pressure_level, indoor_relative_humidity, illuminance, pm2.5, indoor_co2
Room 4	13536	13	voc, sound_pressure_level, indoor_relative_humidity, illuminance, pm2.5, indoor_co2
Room 5	13536	15	voc, sound_pressure_level, indoor_relative_humidity, illuminance, pm2.5, indoor_co2
		2580	supply_air_flow and damper_position

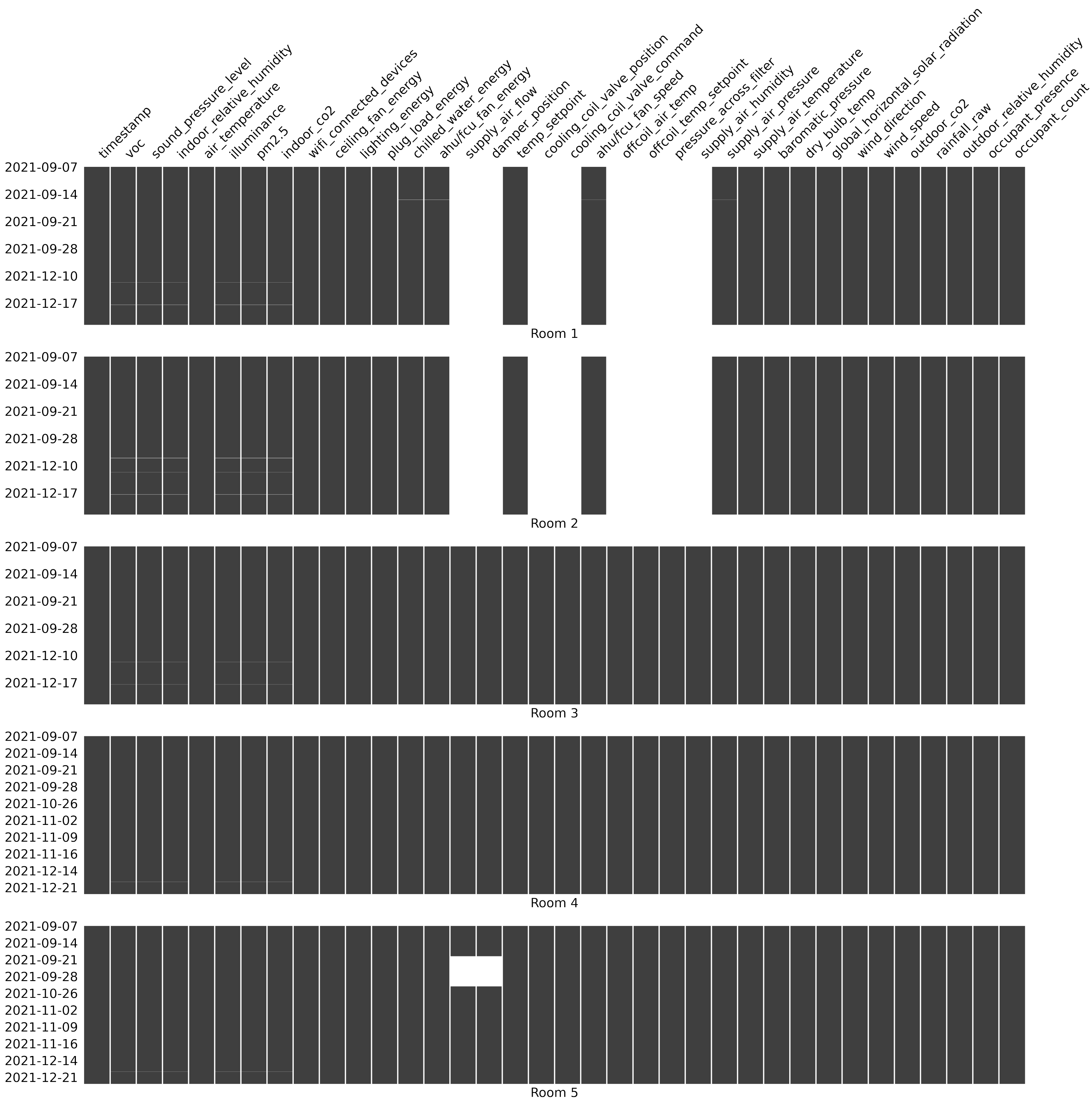

Figure 2. The amount of missing data in each column and their temporal relationship for each room.

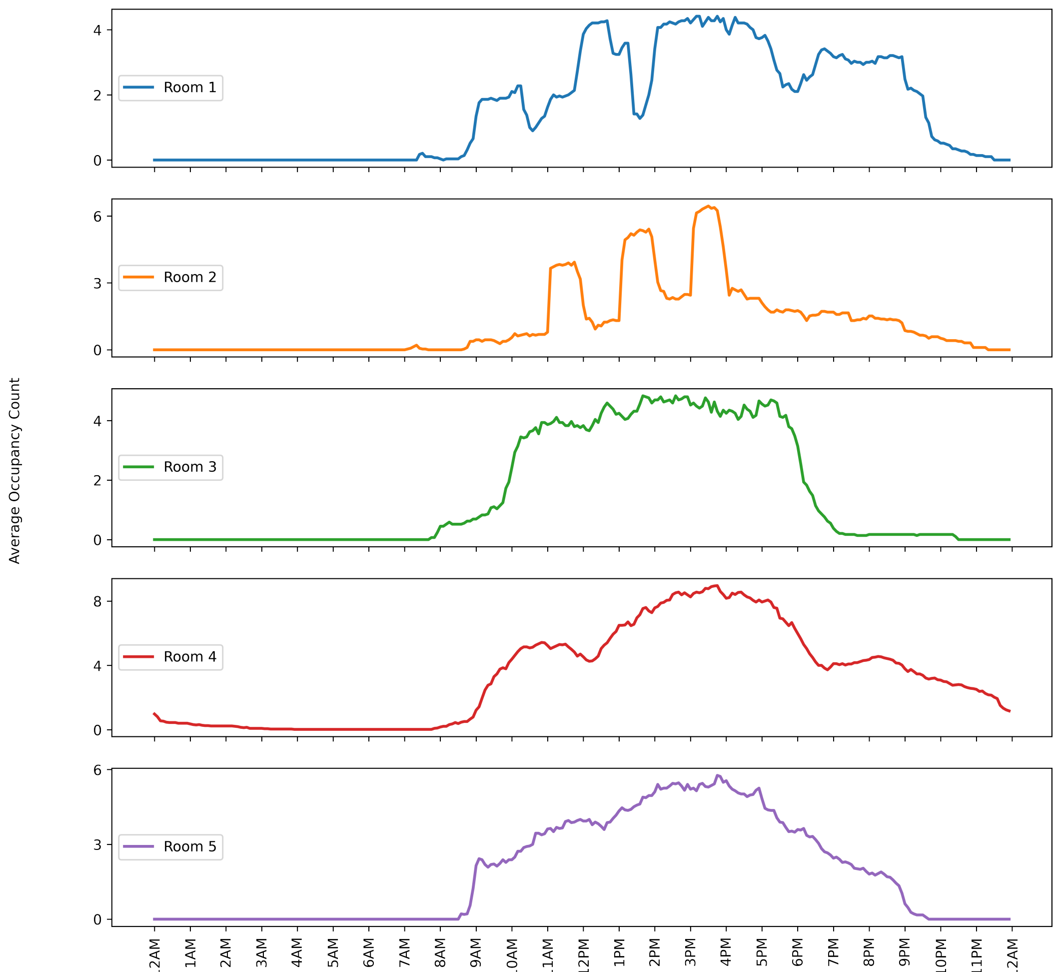

Figure 3. Average occupant count for each room on an average day.

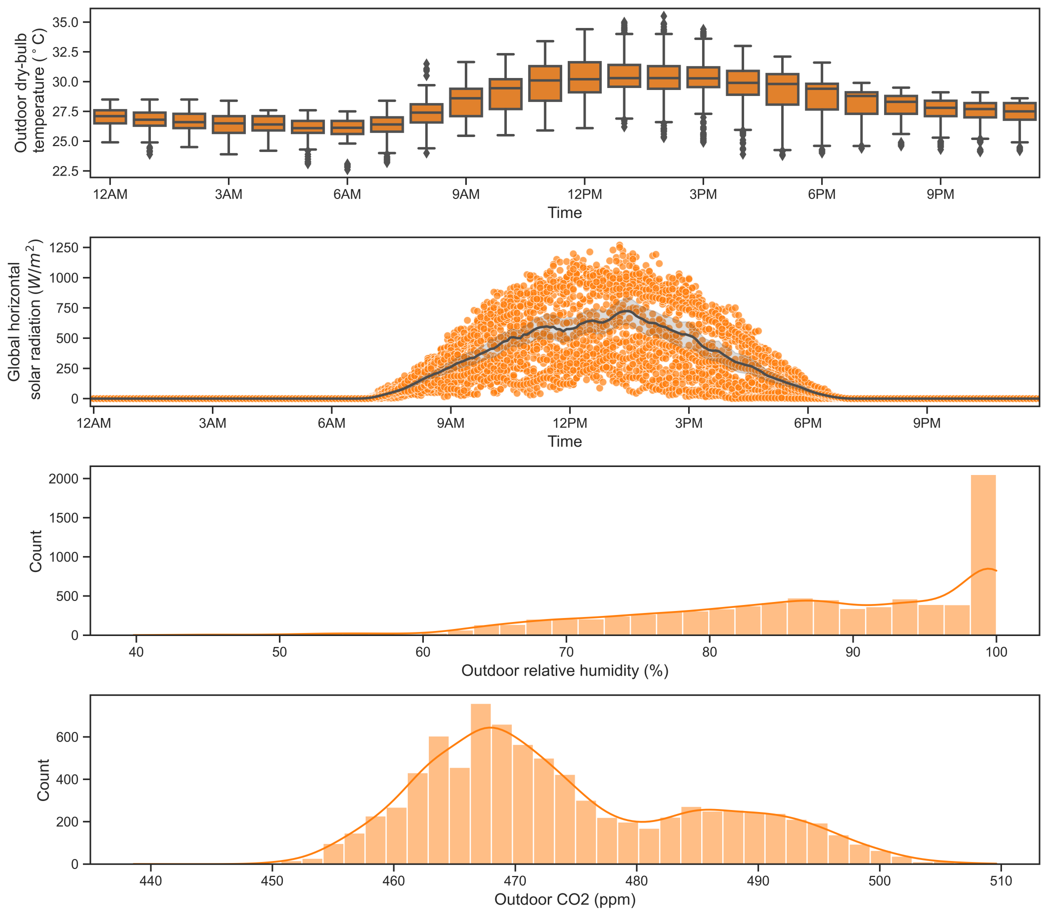

Figure 4. Data visualisations for outdoor dry-bulb temperature, relative humidity and CO_2 levels.

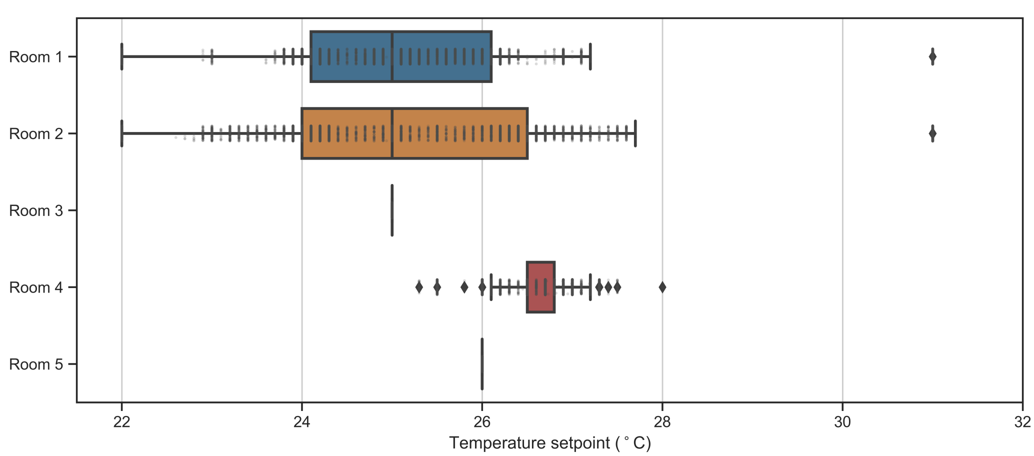

Figure 5. Distributions of room temperature setpoints for each room.

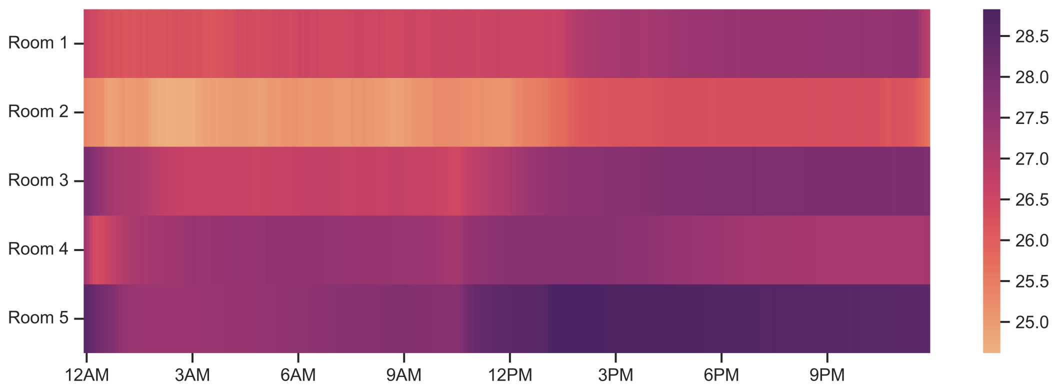

Figure 6. Average room air temperature (°C) for each room.

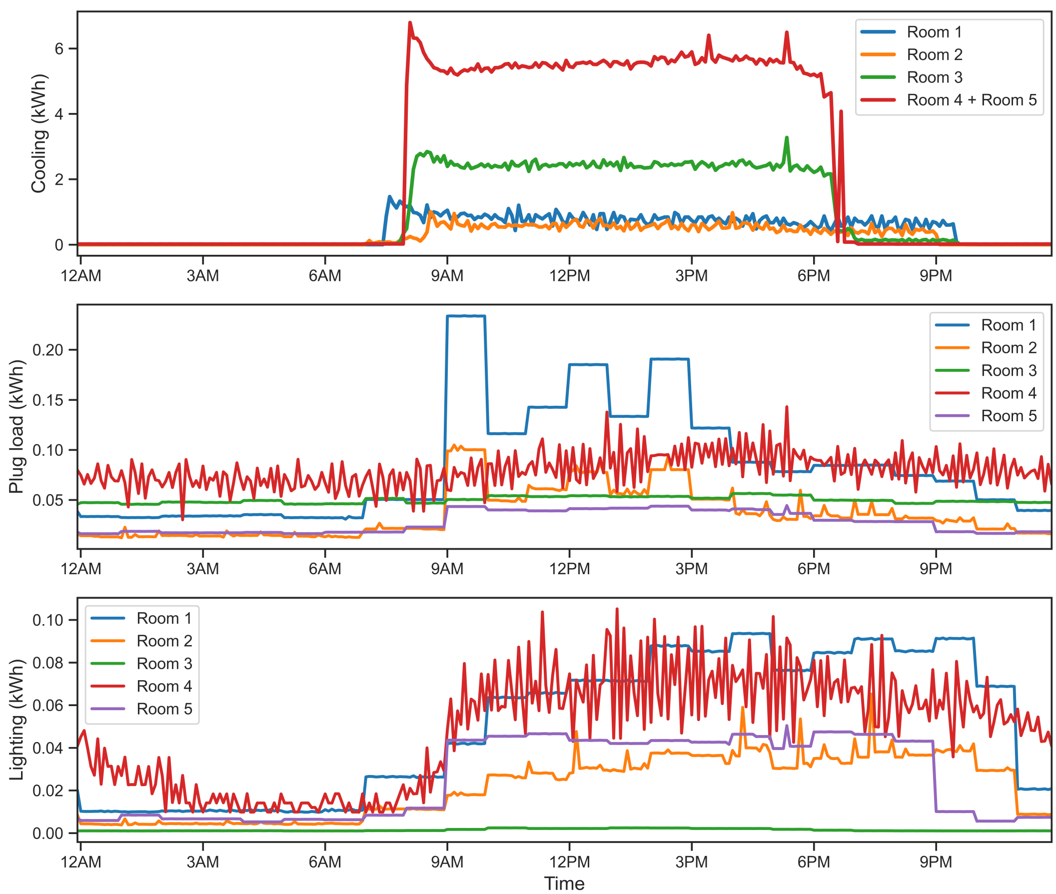

Figure 7. Average daily energy consumption of space cooling, plug load, and lighting for each room. For cooling, it should be noted that Room 3, Room 4 and Room 5 are conditioned by AHUs, which are connected to multiple rooms. In this case, Room 4 and Room 5 share the same AHU and therefore have the same cooling consumption as reflected in the figure.