BUILD SIMUL (2022) 15: 2127–2137
https://doi.org/10.1007/s12273-022-0925-9

ROBOD, room-level occupancy and building operation dataset

Zeynep Duygu Tekler, Eikichi Ono, Yuzhen Peng, Sicheng Zhan, Bertrand Lasternas, Adrian Chong (✉)

Department of the Built Environment, National University of Singapore, 4 Architecture Drive, 117566, Singapore

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 supporting data-driven approaches in the field of occupancy prediction and occupant behaviour modelling, building simulation and control, energy forecasting and various building analytics.

Keywords

building operation data;
occupancy data;
sensor fusion;
occupancy prediction;
building simulation and control

Article History

Received: 11 May 2022
Revised: 12 July 2022
Accepted: 30 July 2022

1 Introduction

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 (IEA 2020). 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 (Orouji et al. 2021), architectural design (Ataman and Dino 2021), environmental conditions (Aydin and Jakubiec 2018), occupancy and occupant interactions with building systems (Tekler et al. 2019a; Peng et al. 2019; Ouf et al. 2021). 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 (Tang et al. 2021; Ding et al. 2022). For instance, the availability of various building data is necessary for physics-based energy models to define model assumptions and inform model calibration (Chong et al. 2021). At the same time, data-driven or machine learning-based models require a sufficient amount of training data to produce reliable prediction results (Peng et al. 2018).

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 (Tekler et al. 2020a). Despite these challenges, there has been a sustained effort within the

E-mail: adrian.chong@nus.edu.sg

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 (Miller et al. 2020) which contains the energy metering data for 1,636 non-residential buildings; BLOND (Kriechbaumer and Jacobsen 2018) an energy consumption dataset for appliances in an office building; flEECe (Paige et al. 2019) an energy use and occupant behaviour dataset for residential buildings; CU-BEMS (Pipattanasomporn et al. 2020), 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 (Schwee et al. 2019; Tekler et al. 2020b; Li et al. 2021). Finally, the latest release of the ASHRAE Global Occupant Behaviour database provides a large compilation of different survey-based and in-situ-based datasets collected from multiple countries and covering various building types both in the commercial and residential sectors (Dong et al. 2022).

In this paper, we release ROBOD, a **R**oom-level **O**ccupancy and **B**uilding **O**peration **D**ataset. 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. This dataset complements the existing ASHRAE Global Occupant Behaviour database by providing a more comprehensive set of data categories for different space types when compared to existing datasets and provide researchers with a rare and unique view into the operations of a net-zero energy building. 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.

2 Methods

2.1 Building and 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 (NUS). SDE4 is a 6-story academic building spanning 8,588 square meters and is accessible 24 hours every day. 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. This certification awarded by the International Living Future Institute (ILFI) recognises green buildings that are able to satisfy 100% of its annual energy needs via on-site renewable energy sources. The room occupancy and building operation data for five rooms was collected as part of this study and the rooms are 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). Specific rooms such as Room 1, Room 2 and Room 5 are open for all university students and staff, and are not limited to those situated in the SDE4 building. On the other hand, Room 3 and Room 4 are only accessible by dedicated administrative staff and researchers that have an assigned seating in these rooms as an access card is required for entry.

The detailed description of each room is provided in Table 1.

2.2 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

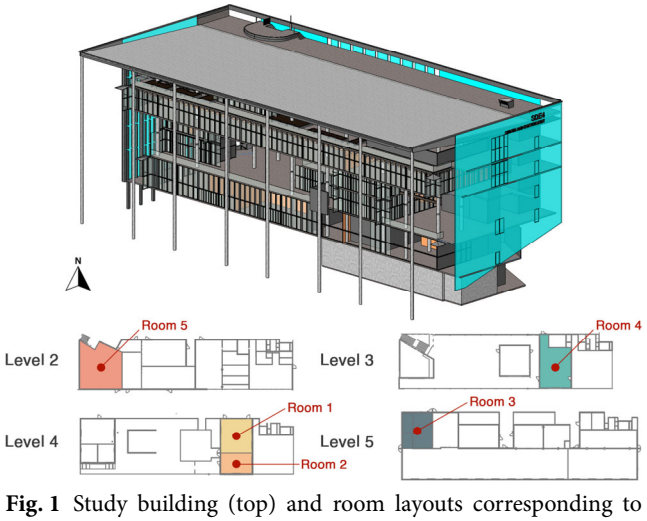


Fig. 1 Study building (top) and room layouts corresponding to the building levels (bottom)

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 occupant 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

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.

2.2.1 *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, $PM_{2.5}$ (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 with their corresponding layouts.

2.2.2 *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, the number of Wi-Fi connected devices are generally higher than the number of occupants in the room as the data recorded does not differentiate between mobile and stationary devices. To use the number of Wi-Fi connected devices to estimate the number of occupants in the room, several filtering steps could be introduced to differentiate between the stationary and mobile devices, before inferring the occupant count based on the number of mobile devices indirectly. These filtering steps could be performed in the case where surveillance cameras are not available to determine the number of occupants in the room. 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.

2.2.3 *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.

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. All lighting and plug load units are manually controlled by occupants. 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,

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 unit (Awair Omni)	0–60000	±10%
	Sound pressure	dB(A)		Unspecified	Unspecified
	Relative humidity	%RH		0–100	±2%RH
	Air temperature	°C		−40 to 125	±0.2 °C
	Illuminance	lux		0–64000	Unspecified
	$PM_{2.5}$	$\mu g/m^3$		0–1000	±15 $\mu g/m^3$
	CO_2	ppm		400–5000	±75 ppm
Wi-Fi	Wi-Fi connected devices	Number	Wi-Fi router (Cisco)	NA[a]	NA[a]
Energy	Ceiling fan energy Lighting energy Plug load energy	kWh	Energy meter (Schneider Electric Acti9 iEM3000)	0–999999	±1%
	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 airflow[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] Cooling coil valve command[b]	%	Valve (Johnson Controls)	0–100	NA[a]
	AHU/FCU fan speed	Hz	Variable speed drive (ABB)	0–50	±0.2%
	Offcoil air temperature[b]	°C	NTC thermistor (Greystone TSDC)	−40 to 60	±0.2°C
	Offcoil temperature setpoint[b]	°C	NA[a]	NA[a]	NA[a]
	Supply air humidity[b]	%RH	Capacitive (Greystone HSDT)	0–100	±2%RH
	Pressure across filter[b] Supply air pressure	Pa	Capacitive (Setra 264)	Unspecified	±1%
	Supply air temperature	°C	NTC thermistor (Greystone TSAP)	−40 to 60	±0.2°C
Outdoor weather[c]	Barometric pressure	hPa	Piezoresistive	600–1100	±0.5hPa
	Dry bulb temperature	°C	Pt100	−40 to 60	±15 °C
	Global horizontal solar radiation	W/m^2	Thermophile	0–2000	2nd class pyranometer
	Wind direction	Degree	Ultrasonic	0–360	±15 RMSE
	Wind speed	m/s		0–60	±0.2 m/s
	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 (Xeron Vision)	NA[a]	NA[a]
	Occupant count	Number			

[a] NA refers to "Not Applicable".

[b] Indicated measurements are not applicable for Room 1 and 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) with corresponding layouts

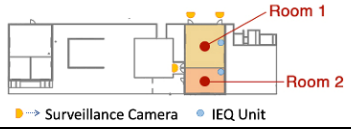
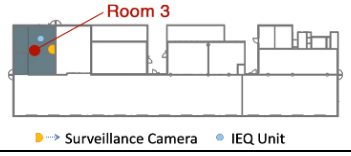
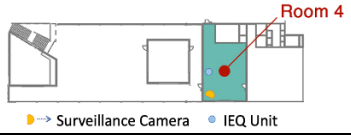
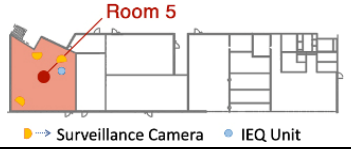
Room	Surveillance camera	IEQ unit	Layout
Room 1	Two surveillance cameras outside of two doors	Mounted vertically to an east side column of the wall	
Room 2	One surveillance camera outside of the door	Mounted vertically to an east side column of the wall	
Room 3	One surveillance camera inside the room	Mounted vertically to an east side column of the wall	
Room 4	One surveillance camera inside the room	Mounted vertically to a west side column of the wall	
Room 5	Three surveillance cameras inside the room	Mounted vertically to an east side column of the 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

Room 3 and Room 4 contain different number of monitors, laptops, desktops, and printers; while Room 5 contains mostly laptops and printers.

2.2.4 *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. Furthermore, there is no operable shading or windows in the study rooms, except for the operable windows in Room 4 which are rarely open based on our observations. Therefore, the state of the windows and shades' impact on HVAC operations is not considered in the dataset.

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 rate at 3,375 and 2,025 cubic meter per hour (CMH), respectively. Variable speed drive (VSD) fans are also used to regulate supply airflow in the rooms to maintain room temperature. The HVAC operating hours for these rooms are set at 07:30 to 21:40.
- Room 3 has dedicated VAV to supplying airflow rate at 900 CMH to maintain room the temperature. It is air-conditioned by an AHU with a supply airflow rate of 1,3470 CMH, serving five other rooms in the building. The HVAC operating hours for this room are set at 08:30 to 18:40.
- Room 4 and Room 5 have dedicated VAV supplying airflow rate at 3,192 CMH and 1,944 CMH to maintain the room temperature. 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. The HVAC operating hours for these rooms are set at 08:30 to 18:40.

2.2.5 *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. The angle definition for wind direction is set in the clockwise direction, where "0°" indicates the north direction. It should be noted that all rooms use the same local weather station and therefore the corresponding measurements for weather data are identical for all rooms.

2.2.6 *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. Due to the use of a passive monitoring approach to monitor occupancy within the study room, the impact of the Hawthorne effect is minimised as compared to the adoption of active monitoring approaches. These approaches often require occupants to carry around wearable sensors or install mobile applications on their smartphone devices to track their locations, which may cause them to change their behaviours or regular routines (Tekler et al. 2019b). 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 to protect the occupants' privacy. The protocols for the data collection has been approved by the host university's Institutional Review Board (NUS-IRB-2021-31).

2.3 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.

3 Data records

ROBOD 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 7 September and 23 December 2021, where the sensor data collected during the weekends were excluded. Several notable holiday periods that occurred during the data collection period includes a public holiday on 4 November 2021 as well as the university's semester break, which occurred between 5 December and 23 December. The chosen period allowed us to capture the changes in occupancy patterns and the building's operation both during the regular semester and the semester break. 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.

4 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.

4.1 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 amount of missing data is represented in terms of the number of rows in the dataset as well as their corresponding percentages (%). The temporal relationship of the missing data is also presented in Figure 2. It should be reiterated that the datasets for Rooms 1 and 2 do not contain columns related to VAV (i.e., Supply airflow, Damper position, Cooling coil valve position and command, Offcoil temperature setpoint,

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 (0.1%)	supply_air_pressure and ahu_fan_speed
		10 (0.1%)	chilled_water_energy and ahu_fan_energy
		14 (0.1%)	voc, sound_pressure_level, indoor_relative_humidity, illuminance, pm2.5, indoor_co2
Room 2	8352	30 (0.3%)	voc, sound_pressure_level, indoor_relative_humidity, illuminance, pm2.5, indoor_co2
Room 3	8352	13 (0.1%)	voc, sound_pressure_level, indoor_relative_humidity, illuminance, pm2.5, indoor_co2
Room 4	13536	13 (0.1%)	voc, sound_pressure_level, indoor_relative_humidity, illuminance, pm2.5, indoor_co2
Room 5	13536	15 (0.1%)	voc, sound_pressure_level, indoor_relative_humidity, illuminance, pm2.5, indoor_co2
		2580 (19.0%)	supply_air_flow and damper_position

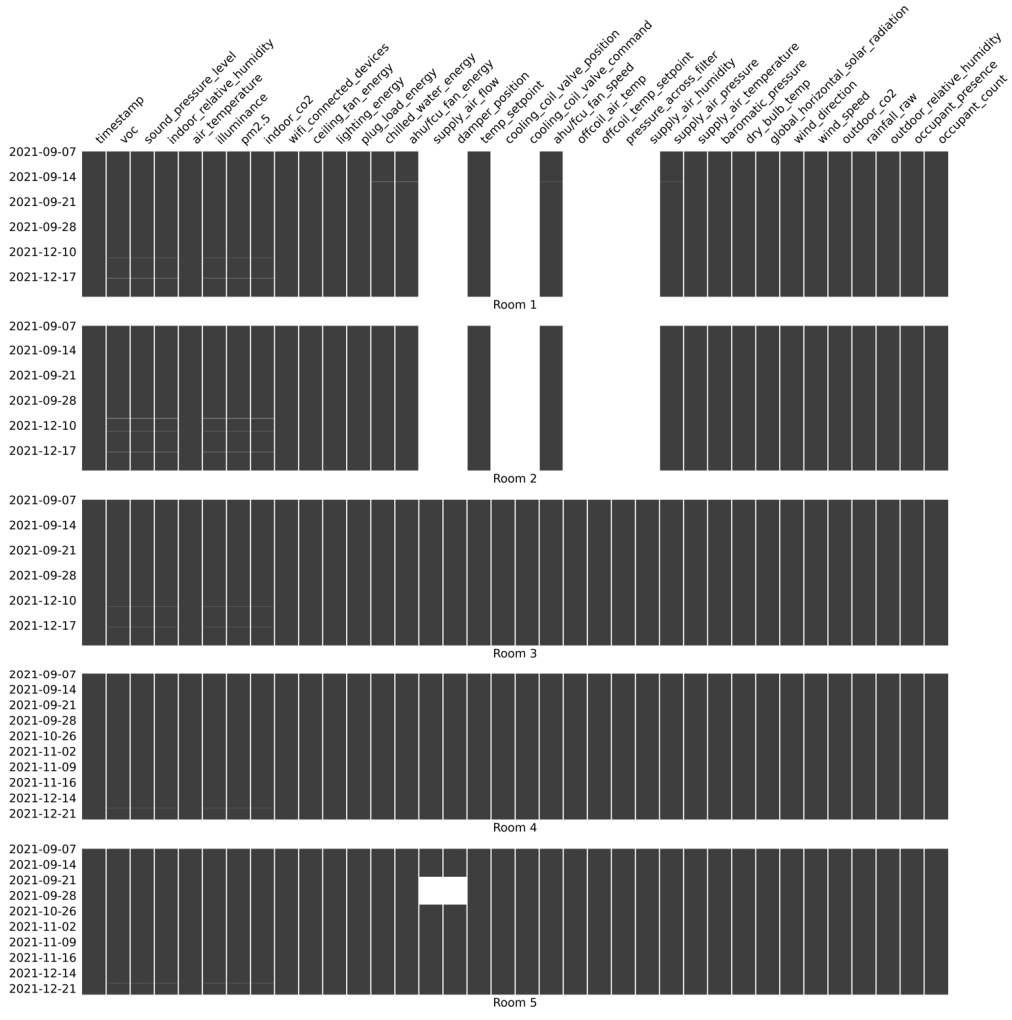


Fig. 2 The amount of missing data in each column and their temporal relationship for each room

Offcoil air temperature, Pressure across filter, and Supply air humidity) as they are conditioned by FCUs, therefore they are not included in the dataset.

4.2 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 a.m., 1 p.m., and 3 p.m., which can be explained by the block lectures that are regularly scheduled during these periods. Room 3 presents

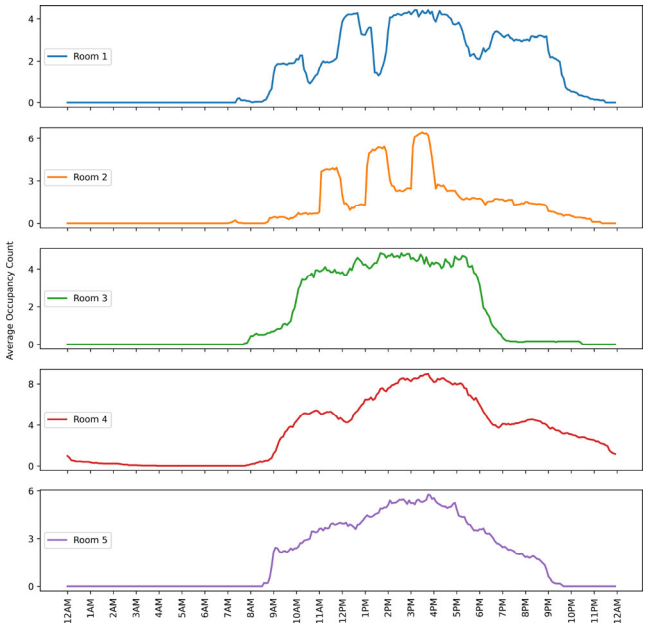


Fig. 3 Average occupant count for each room on an average day

a regular office schedule with the office workers arriving at work between 8 and 10 a.m. and leaving the office at the end of the workday between 6 and 8 p.m. 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 a.m., followed by a sharp drop back to zero at 9 p.m. every day, corresponding with the operational hours of the library space.

4.3 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 p.m. to 4 p.m.). The global horizontal solar radiation can reach over 1000 W/m^2 between 11 p.m. and 3 p.m. 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 airflow 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 and 510 ppm, which is used as the basis of maintaining the indoor CO_2 levels at a standard or comfortable range.

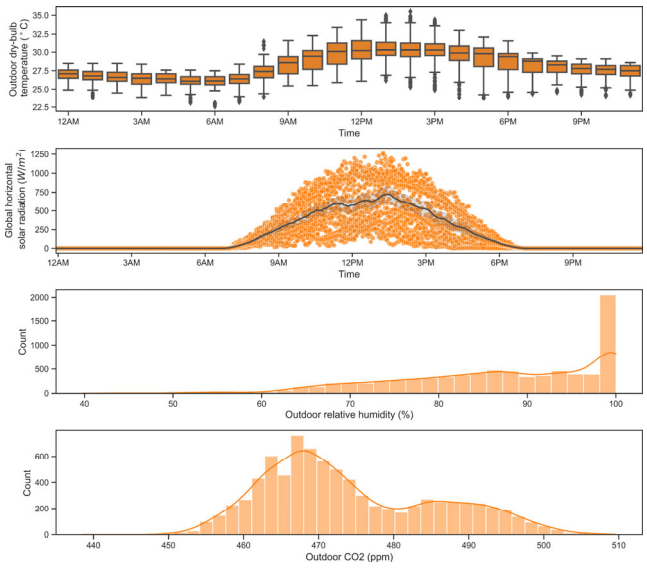


Fig. 4 Data visualisations for outdoor dry-bulb temperature, relative humidity and CO_2 levels

4.4 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.

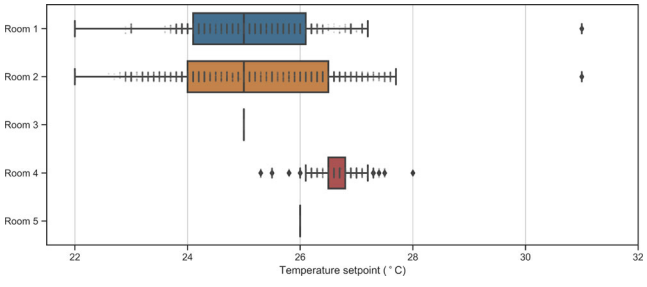


Fig. 5 Distributions of room temperature setpoints for each room

4.5 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 temperature in Room 1, Room 2, and Room 3 tends to be

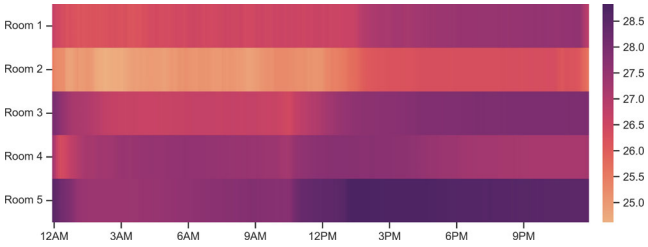


Fig. 6 Average room air temperature (°C) for each room

lower than that in Room 4 and Room 5. Moreover, the air temperature in the afternoon is also higher than that in the morning for all five rooms.

4.6 Energy consumption

Figure 7 summarizes energy consumption of space cooling, plug load, and lighting in each room. The cooling energy consumption is a combination of 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 calculated based on the ratio of the VAV supply airflow over the AHU's total supply airflow. 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 Room 4 and Room 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

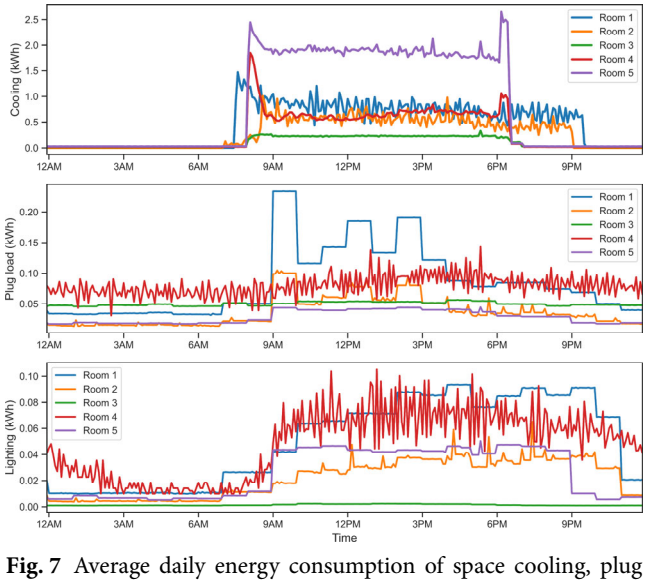


Fig. 7 Average daily energy consumption of space cooling, plug load, and lighting for each room. For cooling, it should be noted that Room 3, Room 4, and Room 5 are conditioned by AHUs, which are connected to multiple rooms

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 needs the electricity from electrical sockets when their owners occupy the rooms. For instance, the plug load consumption in five rooms is nearly constant before 7 a.m. Furthermore, this consumption in Room 1, Room 2, and Room 5 increased simultaneously from 9 a.m. 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 a.m. in most of the rooms.

5 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 (Low et al. 2020), a Random Forest-based imputation algorithm (i.e., MissForest (Stekhoven and Buhlmann 2012)) 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.

6 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, which is available in the Electronic Supplementary Material (ESM) in the online version of this paper. The ROBOD is also provided as a supplementary material, which is uploaded in an open data repository (i.e., Figshare).

7 Concluding remarks

The availability of the building's operation data and high-resolution occupancy information has been crucial in supporting the evaluation of existing models and development of new data-driven approaches. To facilitate these research efforts, this paper describes a comprehensive dataset (i.e., ROBOD) containing multiple data categories together with ground truth occupant presence and count information for different spaces types collected from a net zero energy university building. A total of 181 days of data was collected from five rooms at a sampling resolution of 5 minutes and comprises of two different-sized lecture rooms, an office space for administrative staff, an office space for researchers, and a library space accessible to all students.

Through the conduct of this study, we have experienced several challenges along the way and have summarised their proposed solutions below to aid future studies in overcoming these challenges:

- **Minimising Hawthorne effect**: The use of surveillance cameras over active monitoring approaches (i.e., wearable sensors and smartphone devices) was chosen to passively monitor the users' presence information to minimise the study's impact on the users' behaviours and regular routines.
- **Protecting user privacy**: Personal identifiers were not collected from the occupants at any point during the data collection process to protect their privacy.
- **Sensor failure**: There were specific days during the data collection period where several sensors were not working correctly for certain rooms. As a result, this led to the data collected during these periods being dropped from the final dataset.
- **Missing data**: Due to the presence of a small number of missing data records caused by intermittent sensor failure, an imputation algorithm based on the "missingpy" library was proposed to impute the missing data values.
- **Merging different data categories**: To perform a successful merge of the different data categories, several data pre-processing steps were taken. These steps involve first down-sampling/up-sampling the data records from each data category to follow a 5-minute sampling interval before standardising the timestamp information for each data category to follow the same ISO 8601 date-time format.

Through the availability of ROBOD, we hope to provide researchers with a rare and unique view into the operations of a net-zero energy building, and a useful benchmark against existing buildings. With the continued advancements in construction technology and renewable energy systems, net zero energy buildings will become increasingly common in the future, thereby allowing this dataset to remain relevant.

Finally, there are also future plans to continue extending this data collection effort by increasing the duration of data collection period, as well as including more space types and data categories to further contribute to current research efforts.

Electronic Supplementary Material (ESM): A step-by-step guide is available in the online version of this article at https://doi.org/10.1007/s12273-022-0925-9.

Data availability

The ROBOD is available at
https://doi.org/10.6084/m9.figshare.19234530.v7

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.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Ethical approval

This study does not contain any studies with human or animal subjects performed by any of the authors.

Author contribution statement

All authors contributed to the study conception and design. Material preparation, data collection and analysis, visualisations were performed by Zeynep Duygu Tekler, Eikichi Ono, Yuzhen Peng and Sicheng Zhan. Administrative and technical support during the data collection have been provided by Bertrand Lasternas. Supervision, funding

acquisition and critical feedback during data collection have been provided by Adrian Chong. The first draft of the manuscript was written by Zeynep Duygu Tekler and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

References

Ataman C, Dino İG (2021). Performative design processes in architectural practices in Turkey: Architects' perception. *Architectural Engineering and Design Managemen*, https://doi.org/10.1080/17452007.2021.1995315

Aydin EE, Jakubiec JA (2018). Sensitivity Analysis of Sustainable Urban Design Parameters: Thermal Comfort, Urban Heat Island, Energy, Daylight, and Ventilation in Singapore. In: Proceedings of Building Simulation and Optimization, Cambridge, UK.

Chong A, Gu Y, Jia H (2021). Calibrating building energy simulation models: A review of the basics to guide future work. *Energy and Buildings*, 253: 111533.

Ding Y, Han S, Tian Z, et al. (2022). Review on occupancy detection and prediction in building simulation. *Building Simulation*, 15: 333–356.

Dong B, Liu Y, Mu W, et al. (2022). A global building occupant behavior database. *Scientific Data*, 9: 369.

IEA (2020). IEA Tracking Buildings 2020.

Kriechbaumer T, Jacobsen H-A (2018). BLOND, a building-level office environment dataset of typical electrical appliances. *Scientific Data*, 5: 180048.

Li H, Wang Z, Hong T (2021). A synthetic building operation dataset. *Scientific Data*, 8: 213.

Low R, Tekler ZD, Cheah L (2020). Predicting commercial vehicle parking duration using generative adversarial multiple imputation networks. *Transportation Research Record*, 2674: 820–831.

Miller C, Kathirgamanathan A, Picchetti B, et al. (2020). The building data genome project 2, energy meter data from the ASHRAE great energy predictor III competition. *Scientific Data*, 7: 368.

Orouji P, Hajian R, Moradi M, et al. (2021). Atlas of heating: Identifying regional climate-dependent heat demands in residential buildings of Iran. *Building Simulation*, 14: 857–869.

Ouf MM, Park JY, Gunay HB (2021). A simulation-based method to investigate occupant-centric controls. *Building Simulation*, 14: 1017–1030.

Paige F, Agee P, Jazizadeh F (2019). flEECe, an energy use and occupant behavior dataset for net-zero energy affordable senior residential buildings. *Scientific Data*, 6: 291.

Peng Y, Rysanek A, Nagy Z, et al. (2018). Using machine learning techniques for occupancy-prediction-based cooling control in office buildings. *Applied Energy*, 211: 1343–1358.

Peng Y, Nagy Z, Schlüter A (2019). Temperature-preference learning with neural networks for occupant-centric building indoor climate controls. *Building and Environment*, 154: 296–308.

Pipattanasomporn M, Chitalia G, Songsiri J, et al. (2020). CU-BEMS, smart building electricity consumption and indoor environmental sensor datasets. *Scientific Data*, 7: 241.

Schwee JH, Johansen A, Jørgensen BN, et al (2019). Room-level occupant counts and environmental quality from heterogeneous sensing modalities in a smart building. *Scientific Data*, 6: 287.

Stekhoven DJ, Buhlmann P (2012). MissForest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28: 112–118.

Tang R, Wang S, Sun S (2021). Impacts of technology-guided occupant behavior on air-conditioning system control and building energy use. *Building Simulation*, 14: 209–217.

Tekler ZD, Low R, Blessing L (2019a). Using smart technologies to identify occupancy and plug-in appliance interaction patterns in an office environment. *IOP Conference Series: Materials Science and Engineering*, 609: 062010.

Tekler ZD, Low R, Blessing L (2019b). An alternative approach to monitor occupancy using bluetooth low energy technology in an office environment. *Journal of Physics: Conference Series*, 1343: 012116.

Tekler ZD, Low R, Gunay B, et al. (2020a). A scalable Bluetooth Low Energy approach to identify occupancy patterns and profiles in office spaces. *Building and Environment*, 171: 106681.

Tekler ZD, Low R, Zhou Y, et al. (2020b). Near-real-time plug load identification using low-frequency power data in office spaces: Experiments and applications. *Applied Energy*, 275: 115391.