Investigating the predictors of air conditioning use in residential buildings by comparing Single- and Multi-Domain GLMM approaches

Arianna Latini [1[0000-0003-1707-1991]], Elisa Di Giuseppe [1[0000-0003-2073-1030]], Gabriele Bernardini[1[0000-0002-7381-4537]], Andrea Gianangeli[1[0000-0003-2936-437X]], Marco D'Orazio[1[0000-0003-3779-4361]]

[1] Construction, Civil Engineering and Architecture Department, Università Politecnica delle Marche, 60131 Ancona, Italy
* a.latini@staff.univpm.it

Abstract. Climate change is causing an increasing cooling demand in residential buildings. Understanding the drivers behind occupants' use of air conditioning is critical for accurate building energy performance analysis. However, occupant-building interactions are highly variable and influenced by multidisciplinary factors, which cause critical uncertainty in behavioural modelling and energy use prediction. This study proposes the use of GLMMs to investigate if the inclusion of multi-domain factors (including physical, behavioural, and contextual domains) in behavioural models increase the predictive performance, in comparison with single-domain models. Results from a monitoring campaign in three residential building apartments reveal a better performance of the multi-domain model in predicting occupant behaviour. Insights obtained from the multi-domain model then reveal that daily variability and apartment differences significantly influence air conditioning status. Occupancy, outdoor humidity, and CO_2 levels increase the likelihood of activation, while high air temperature differences between indoors and outdoors, high indoor humidity and window opening reduce it.

Keywords: Multi-domain, Residential Building, Occupant Behaviour, Indoor environmental quality, Human-building interaction

1 Introduction

Occupants of residential buildings are responsible for a significant amount of energy consumption [1]. A special concern is devoted to cooling energy demand which is rapidly increasing in response to climate change and comfort expectations. Understanding the drivers of occupants' interaction with air conditioning systems is crucial for accurate building energy performance analysis [2]. It is well-known that the interaction between occupants and buildings is extremely variable and multidisciplinary (i.e., involving engineering, psychology, environmental sciences), with the complex interplay of environmental (representing the conditions of the indoor environment and its quality,

e.g. temperature, humidity, Co2 concentration), personal (such as age, gender, culture), and contextual (e.g. building typology and intended use, hour of the day also in reference to different seasons, effective interaction possibilities) variables [3]. This complexity introduces uncertainty in behavioural pattern identification and consequent energy use prediction [2, 4]. In this sense, it becomes essentially to balance complexity (in respect of the type and number of considered variables), generalizability, and representativeness of effective human-building interactions to provide reliable, "detailed enough" predictive models [2, 5]. In this context, the concept of adaptive thermal comfort, which acknowledges the dynamic relationship between occupants and their environment, and their ability to adapt to varying indoor conditions over time [6] offers a valuable theoretical foundation. This concept supports the inclusion of behavioural and contextual variability in modelling approaches, further reinforcing the need for integrated, multi-domain analyses in occupant behaviour research. Given this context, this study aimed to verify if and how the inclusion of multi-domain factors (which comprise physical, behavioural, and context variables) enhances the predictive performance of occupant behavioural models compared to single-domain models (only based on physical variables). In this way, this paper contributes to exploring how more complex approaches could include relevant parameters in respect to easy to apply (but too simplified) single-domain models, and thus define the first steps towards a more "conscious" adoption of data-driven techniques. This objective is pursued by the development of Generalized Linear Mixed Models (GLMMs) and subsequent testing on high-resolution sensor data collected from multiple residential apartment buildings during the summer season. Those models offer a balanced method between interpretability and flexibility to capture key predictors of occupant behaviour with high explanatory power considering behavioural diversity [7, 8]. These models are a complex extension of commonly used techniques, such as logistic regression, widely applied for thermostat adjustment predictions [2], by integrating hierarchical data structure into binary outcome variables, such as the on/off status of air conditioning systems.

2 Literature review

For decades, occupants' behaviours have been modelled using a variety of approaches (e.g., deterministic and statistical methods, and artificial intelligence). The common thread is the need to properly capture the behavioural dynamics, to construct predictive models that integrate observed data with the most influencing drivers [2]. In literature, many behavioural models still rely on single variables at a time (i.e., indoor air temperature, CO2 concentrations) [2, 5] (single-domain models), which limits their explanatory power. In addition, most studies were conducted in offices, while other building typologies, like residential buildings, are currently understudied [2, 9], especially in relation to air conditioning behaviour [10]. As a consequence, robust computational approaches integrating multi-domain factors (including physical, behavioural, and contextual domains) are required to identify the main influential factors of occupant behaviour and improve the ability to predict interaction with building systems. Nowadays, Machine Learning (ML) techniques are increasingly employed for advancing occupant

behaviour research, due to their high predictive accuracy and ability to capture non-linear relationships [2]. However, the application of these data-driven approaches (which can be assumed as "black box" systems) should be supported by the identification of key variables inputs, "to achieve a reliable prediction" [4]. In fact, the possibility of a lack of interpretability could be problematic in research fields where understanding the cause-effect relationship is essential, because it limits the ability of researchers to understand underlying behavioural mechanisms. Moreover, ML models require large datasets [8], which are often limited in residential buildings, due to privacy issues and data collection challenges over long periods, thus reducing generalizability.

3 Method

3.1 Residential building data collection

A field campaign was carried out in three apartments (between 49-87 m^2, 2 occupants each) within two multi-story social housing buildings in Reggio Emilia (Italy), which are the demonstrators of the LIFE SUPERHERO project [11]. Details about the building envelope and energy performance can be found in a previous research paper by the authors [12]. A physical monitoring campaign took place from July to September 2022 before the buildings were retrofitted with the installation of external thermal insulation and the replacement of windows. A 10-minute timestamp was set to acquire the following data:

- Indoor environmental conditions are recorded by small wall-mounted sensors [13] in living rooms and bedrooms, considering air temperature (range: 0-50°C), relative humidity (range: 0-100%), CO2 concentration (up to 10,000 ppm), occupants' presence (Boolean, 0/1 when the room is empty/occupied);
- Windows status by using binary state sensors (Boolean, 0/1 when windows are closed/opened);
- Air conditioning status recorded by an energy counter which data were converted in Boolean values 0/1 when AC is off/on depending on the consumed energy at each timestamp [14, 15];
- Outdoor environmental conditions are measured by a weather station located on the roof of one of the buildings [16], considering air temperature (range: -40 to 65°C), relative humidity (range: 1-100%), solar radiation (range: 0-1800W/m^2), wind speed (range: 1-322km/h), wind direction, precipitation (up to 1000 mm/h).

3.2 GLMM development

The hypothesis is that the inclusion of multi-domain factors, namely physical, behavioural, and context variables, significantly enhances the predictive performance of occupant behavioural models. Thus, a Generalized Linear Mixed Model (GLMM) was computed using a binomial logit link function. The basic theory of the LMM is that the dependent variable response is the sum of fixed factors, which are the variables of interest monitored during the study, and random factors that can influence the covariance

of the data. In this study, according to literature works [2, 5], we explored each of the records in the 10-minute timestamp dataset as single entries, thus not assuming an analysis into time series. The dependent variable was the air conditioning status, a binary indicator of whether the system is on or off at a given time in each monitored room. Indoor and outdoor environmental variables, windows opening status and contextual factors were considered independent variables used as fixed effects (**Table 1**).

Table 1. Metrics considered for GLMMs generation. *continuous variables were normalized according to previous literature methods [2, 5]; ^predicted variable in Section 2.3 models.

Fixed Effects	Variable list	Variable type
Indoor Environment (IE)	Air Temperature	Continuous*
	Relative Humidity	Continuous*
	CO2 concentration	Continuous*
Outdoor Environment (OE)	Air Temperature	Continuous*
	Relative Humidity	Continuous*
	Solar radiation	Continuous*
	Wind speed	Continuous*
	Fair wind	Binary [0,1]
Occupant Behaviour (OB)	AC status^	Binary [0,1]
	Window status	Binary [0,1]
Context (C)	Occupancy	Binary [0,1]
	Day Number	[1 - 48]
	Day Time	[Morning, Afternoon, Evening, Night]

A random intercept varying among apartments and rooms was included in the model concerning the nested random effects (i.e., specific rooms within specific apartments). This structure accounts for variability at the apartment level as well as within different zones of the same apartment. In addition, a random intercept for each day of the monitoring period was incorporated into the model to consider any potential autocorrelation in the data across the monitoring period. The general specification of the model was as follows: *Dependent Variable ~ Fixed Effects Variables + (1 | Apartment/Room) + (1 | + DayNumber)*

3.3 GLMM testing

Two models of different complexity were developed by adding subsequent factors (**Table 2**), with model#1 being a single-domain model and model#2 a multi-domain model. The Variance Inflation Factor (VIF) was computed to diagnose collinearity between predictors in each model. Thus, these indices measure how much the variance of the regression coefficient estimate is inflated due to the correlation between that predictor and the others. To keep those scores lower than the threshold value equal to 5

[17], some predictors (Indoor Air Temperature and Outdoor Air Temperature) were combined by computing their difference value to deal with collinearity (ΔT).

Table 2. Proposed models, predictors and domain

Model	Predictors	Domains
1	AC ~ IE + OE	Single-domain: Physical (Thermal, Air Quality)
2	AC ~ IE + OE + OB + C	Multi-domain: Physical (Thermal, Air Quality), Behavioural, Context

Then, the following performance metrics were computed and compared across the two models to select the best performing GLMM:

- Akaike Information Criterion (AIC): lower values indicate better model fit;
- Bayesian Information Criterion (BIC): similar to AIC, but balanced with model complexity to support model parsimony;
- R^2 Marginal and R^2 Conditional to extract the proportion of variance explained by fixed effects and by both fixed and random effects, respectively [18]
- ANOVA (model#1 vs model #2) to test whether adding multi-domain predictors significantly improves the model fit. P-values < 0.05 were considered significant thus demonstrating that the inclusion of behavioural and context variables improves the predictive performance of the GLMM.

4 Results

The sample size dataset was composed of 56642 observations, collected in 7 rooms (i.e., 3 living rooms, 4 bedrooms), from July 29th and September 14th, 2022 (i.e., 48 monitoring days). The statistical analysis was carried out using the statistical software R [19].

4.1 Raw data analysis

Fig. 1 shows the analysis of AC activation frequency by apartment and time of day. In general, nighttime shows the highest frequency of activation (34% of the total), while afternoon and evening exhibit a comparable AC demand (25-26%). Apartment#3 presented the most consistent activation frequency, especially during the evening (54%) and night (60%). While apartment#1 presents a gradual daily increase, peaking in the evening (52%), apartment#2 is the apartment with the lowest activation frequency with no time period exceeding 17%. **Fig. 2** illustrates the frequency of apartments being occupied, according to the time of the day. In general, morning has the lowest occupancy, possibly due to occupants leaving for work/school. Afternoon and evening show a moderate occupancy frequency in all apartments. As expected, nighttime is the most occupied period, but the frequency may be underestimated, as the PIR sensor only activates in the case of users' movements, thus not registering presence with asleep tenants, leading to lower occupancy counts.

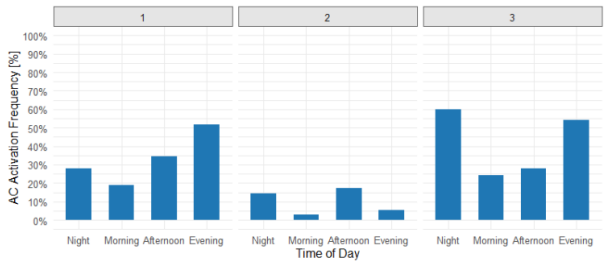

Fig. 1. Frequency of AC activation across Apartments and Day Time

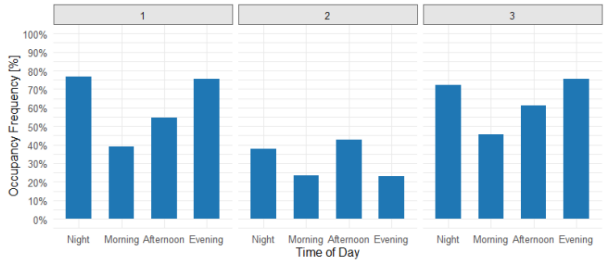

Fig. 2. Frequency of occupancy period across Apartments and Day Time

Table 3 reports the descriptive statistics on indoor and outdoor environmental conditions during the occupancy period, when the air conditioning was off and on, respectively. The analysis highlights differences in environmental conditions with lower indoor air temperature and humidity and higher outdoor temperature, humidity and solar radiation, during periods of AC activation supporting the hypothesis that building overheating due to external conditions is a determinant of the decision to turn on the cooling system. In addition, CO_2 concentrations were higher when the AC was on, probably associated with reduced natural ventilation in the presence of windows closed and/or occupants' smoking activity.

Table 3. Descriptive statistics of indoor and outdoor conditions during the occupancy period

Environmental Condition Variables	**AC status = 0 Mean ± sd**	**AC status = 1 Mean ± sd**
T_{indoor} (°C)	27.14 ± 1.07	26.67 ± 1.43
RH_{indoor} (%)	52.50 ± 6.30	41.27 ± 6.21
CO2 (ppm)	618.96 ± 367.97	1043.8 ± 556.85
$T_{outdoor}$ (°C)	25.32 ± 4.84	26.58 ± 5.50
$RH_{outdoor}$ (%)	63.70 ± 16.25	59.95 ± 18.55
SolRad (W/m^2)	197.88 ± 300.92	269.70 ± 311.34

4.2 GLMM selection

Table 4 shows that Model #2, which includes behavioural and contextual variables, outperforms Model #1 in terms of AIC, BIC, and R^2 metrics (marginal and conditional $R^2 > 0.80$). ANOVA confirms that adding these domains significantly improves model fit ($p < 0.05$). Furthermore, multicollinearity issues present in Model #1 (e.g., VIF > 5 for ΔT and outdoor humidity) were reduced in Model #2.

Table 4. GLMM testing and comparison

GLMM	AIC	BIC	R^2_m, R^2_c	Chisq(df)	p-value
#1 AC ~ IE + OE	25076	25175	0.65,0.79		
#2 AC ~ IE + OE + OB + C	14996	15139	0.96,0.97	9939(5)	< 2e-16 ***

4.3 Insight from the selected GLMM

Model #2 highlights the importance of accounting for temporal, apartment, and room-level variability. Daily variability (variance: 1.77±1.33) has the largest effect, followed by apartment-level (0.75±0.86), with minimal room-level differences (0.03±0.17). **Table 5** reveals that shows that all predictors except solar radiation and wind are significant. Occupancy (Std.Coef. = 10.53), outdoor humidity, CO_2, and time of day are the strongest positive predictors. Conversely, negative standardized coefficients indicate that an increase in air temperature difference, indoor humidity, wind speed, windows opening and the time of day (morning) are more likely to reduce the probability of turning the AC on, with indoor humidity and temperature having a greater impact (-2.59 and -1.32).

Table 5. GLMM results. Statistically significant p-values: < 0.000 '***', 0.001 '**', 0.01 '*'

Domains	Fixed Effect	Std.Coef	95% CI	p-value	GVIF
IE	ΔT	-1.32	[-1.51, -1.14]	***	3.95
	Co2 concentration	0.48	[0.43, 0.53]	***	1.09
	Indoor Humidity	-2.59	[-2.68, -2.50]	***	1.08
OE	Solar Radiation	-0.06	[-0.14, 0.02]		1.99
	Outdoor Humidity	1.34	[1.16, 1.52]	***	3.64
	Wind Speed	-0.12	[-0.17, -0.07]	***	1.10
	Fair Wind	0.01	[-0.08, 0.11]		1.01
OB	Window status	-0.65	[2.73, 18.34]	*	1.06
C	Occupancy	10.53	[-0.72, -0.59]	***	1.00
	Day Time_ Morning	-0.89	[-1.03, -0.74]	***	1.34
	Day Time_Afternoon	0.42	[0.22, 0.63]	***	
	Day Time_Evening	0.44	[0.28, 0.59]	**	

The predictors with the strongest positive and negative influence are represented in **Fig. 3** and **Fig. 4**, respectively. During occupancy periods, the correlations indicate that the probability of occupants activating the cooling systems increases in the presence of higher outdoor humidity (green curves) and CO2 concentrations, particularly during the afternoon and evening. In contrast, during the period with closed windows, the lower the indoor humidity (red line), and the lower the indoor air temperature compared to the outdoor temperature, the higher the probability of air conditioning being active.

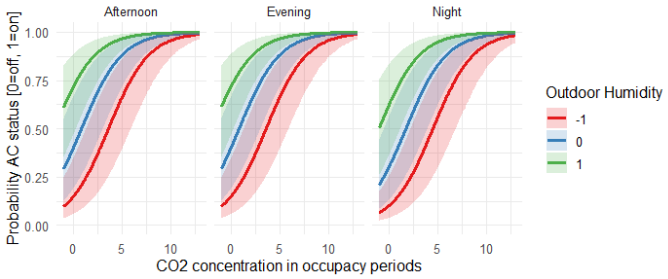

Fig. 3. AC status model prediction (0=off, 1=on) in occupancy period with the most relevant and positively correlated predictors. Actual ranges of normalized x-axis values: CO2 400-6458ppm (in case of occupants smoking near sensors); Outdoor Humidity 23-96%.

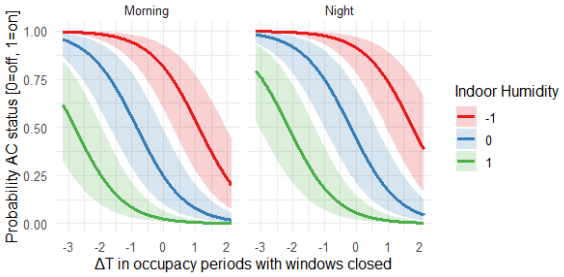

Fig. 4. AC status model prediction in occupancy period with the most relevant and negatively correlated predictors. Actual ranges of normalized x-axis values: ΔT 13.68-11.89°C; Indoor Humidity 27-72%.

5 Discussion

The results confirm expected behavioural patterns: occupancy frequency is closely associated with AC usage, especially during warmer hours of the day. Apartments #1 and #3, which show higher occupancy especially in the evening and night, also present higher AC activation rates. Apartment #2, with generally low occupancy, corresponds to limited AC use. The higher CO_2 concentration during AC use may indicate a reduction in natural ventilation, possibly due to closed windows or indoor smoking,

suggesting a trade-off between cooling and indoor air quality. In further research, model fixed effects will be extended by introducing an additional contextual categorical variable (i.e., Smoking tenants [yes – no]) to better correlate results with occupants' habits. The lower AC usage in the morning could be linked to cooler indoor conditions due to nighttime AC use, particularly when outdoor temperatures exceed indoor ones (negative ΔT). This supports the hypothesis that thermal inertia and nighttime cooling influence morning behaviour. The model's strong performance after incorporating behavioural and contextual variables confirms the value of multi-domain approaches in accurately predicting occupant behaviour in residential environments. The random effects analysis further supports the importance of accounting for temporal and spatial variability to avoid overgeneralization. Finally, the reduction of multicollinearity in Model #2 and the strong explanatory power of occupancy and environmental factors reinforce the need for integrating detailed indoor monitoring with behavioural data to develop more robust predictive models for cooling demand.

6 Conclusion

The outcomes of this paper allow to confirm the hypothesis about the possibility of enhancing the predictive performance of occupants' behavioural models by including multi-domain factors, namely physical, behavioural, and contextual variables. In addition, the proposed GLMM demonstrates the importance of evaluating the influence of multi-domain factors on air conditioning use in residential buildings. Indeed, by including environmental conditions, behavioural and contextual variables, as well as random effects to capture hierarchical structure at apartment and time level, the model provides a solid statistical approach for interpreting the complex and multi-dimensional nature of human-building interaction. The insights obtained from the selected multi-domain occupants' behavioural model serve as the starting point for more advanced behavioural modelling. Indeed, GLMM offers valuable insights that can facilitate the selection process of relevant features to simplify the application of predictive algorithms in data-driven approaches, including those using ML techniques for, thus reducing the risk of overfitting while accounting for cause-effect relationship.

7 Acknowledgement

This work was supported by the Project LIFE SUPERHERO (LIFE19 CCA/IT/001194) "SUstainability and PERrformaces for HEROTILE-based energy efficient roofs", performed with the contribution of the European Union's LIFE Programme "Climate Change Adaptation".

8 References

1. Eurostat (2023) Energy consumption in households. https://ec.europa.eu/eurostat/web/products-eurostat-news/-/ddn-20220617-1. Accessed 4 Apr 2024

2. Mylonas A, Tsangrassoulis A, Pascual J (2024) Modelling occupant behaviour in residential buildings: A systematic literature review. Build Environ 265:111959. https://doi.org/10.1016/j.buildenv.2024.111959
3. Chinazzo G, Andersen RK, Azar E, et al (2022) Quality criteria for multi-domain studies in the indoor environment: Critical review towards research guidelines and recommendations. Build Environ 226:109719. https://doi.org/10.1016/j.buildenv.2022.109719
4. Zhang W, Wu Y, Calautit JK (2022) A review on occupancy prediction through machine learning for enhancing energy efficiency, air quality and thermal comfort in the built environment. Renewable and Sustainable Energy Reviews 167:112704. https://doi.org/10.1016/j.rser.2022.112704
5. Franceschini PB, Schweiker M, Neves LO (2024) Predictive modelling of multi-domain factors on window, door, and fan status in naturally ventilated school classrooms. Build Environ 264:111912. https://doi.org/10.1016/j.buildenv.2024.111912
6. Humphreys M, Nicol F, Roaf S (2015) Adaptive Thermal Comfort: Foundations and Analysis. Routledge
7. Haldi F, Calì D, Andersen RK, et al (2017) Modelling diversity in building occupant behaviour: a novel statistical approach. J Build Perform Simul 10:527–544. https://doi.org/10.1080/19401493.2016.1269245
8. Qiao Q, Yunusa-Kaltungo A, Edwards RE (2021) Towards developing a systematic knowledge trend for building energy consumption prediction. Journal of Building Engineering 35:101967. https://doi.org/10.1016/j.jobe.2020.101967
9. Schweiker M, Ampatzi E, Andargie MS, et al (2020) Review of multi-domain approaches to indoor environmental perception and behaviour. Build Environ 176:106804. https://doi.org/10.1016/j.buildenv.2020.106804
10. Mun SH, Kwak Y, Huh JH (2019) A case-centered behavior analysis and operation prediction of AC use in residential buildings. Energy Build 188–189:137–148. https://doi.org/10.1016/j.enbuild.2019.02.012
11. LIFE SUPERHERO - SUstainability and PERformances for HEROTILE-based energy efficient roofs. https://www.lifesuperhero.eu/. Accessed 14 Nov 2023
12. Latini A, Di Giuseppe E, Bernardini G, et al (2024) A Clustering Method for Identifying Energy-Related Behaviour: The Case-Study of LIFE SUPERHERO Project. In: Proceedings of the 11th International Conference of Ar . Tec . (Scientific Society of Architectural Engineering)
13. ELSYS.se ERS CO2. https://www.elsys.se/en/ers-co2/. Accessed 13 Feb 2024
14. ELSYS.se EMS Door. https://www.elsys.se/en/ems-door/. Accessed 13 Feb 2024
15. Innovation OL ORNO OR-WE-514. https://orno.pl/en/product/1078/1-phase-energy-meter-with-rs-485-100a-rs-485-port-mid-1-module-din-th-35mm. Accessed 13 Feb 2024
16. Instrument D DAVIS Vantage Pro Weather Station. https://www.davisinstruments.com/pages/vantage-pro2. Accessed 13 Feb 2024
17. Hair JF, Hult GT, Ringle C, Sarstedt M (2017) A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)
18. Ferguson CJ (2009) An Effect Size Primer : A Guide for Clinicians and Researchers. 40:532–538. https://doi.org/10.1037/a0015808
19. (2021) R Studio. https://www.rstudio.com. Accessed 31 May 2021