ISBN: 978-628-95207-8-1. ISSN: 2414-6390. Digital Object Identifier: 10.18687/LACCEI2024.1.1.1411

# Occupant behavior and air conditioning usage revealed from sensor fusion applying the k-means clustering method

Erick Reyes[1], Alisson Dodón[1], and Miguel Chen Austin[1,2,3,*]
[1] Grupo de Investigación Energética y Confort en Edificaciones Bioclimáticas (ECEB), Facultad de Ingeniería Mecánica, Universidad Tecnológica de Panamá, Ciudad de Panamá, Panamá, [erick.reyes1, alisson.dodon, miguel.chen]@utp.ac.pa
[2]Centro de Estudios Multidisciplinarios en Ciencias, Ingeniería y Tecnología (CEMCIT-AIP), Ciudad de Panamá, Panamá
[3]Sistema Nacional de Investigación (SNI), Clayton Ciudad de Panamá, Panamá
[*]Corresponding author: miguel.chen@utp.ac.pa

***Abstract*—The knowledge of the occupant's behavior in a building allows the evaluation of the occupant's comfort in it since it takes into consideration aspects of his surroundings or environment that affect them directly, as well as the consideration of entrance/exit in the room and the energy consumption, which allows the evaluation of improvement alternatives in terms of building design. In this study, the k-means algorithm was implemented on data collected (temperature, relative humidity, carbon dioxide) in a room of a two-story residence for one year. The results show that carbon dioxide data is the best for detecting occupant presence, however, all three types of variables were able to detect the use of air conditioning in the case study.**

***Index Terms*—correlation coefficient analysis, environmental data, k-means algorithm, occupancy detection, residential buildings**

## I. Introduction

Occupant behavior in buildings has been a factor of research over the last few years, which provides significant benefits for building space design, heating, air conditioning, lighting and ventilation systems, and improved thermal control. In [1] they employ a cluster analysis of occupancy schedules in the United States in residential buildings from data collected in the American Time Use Survey Data (ATUS) over 12 years (2006-2017). This data was categorized into small groups: age and weekdays, then divided into activities that people do, both "at home" and "away from home" where the presence/non-presence of occupants was mapped in the dwelling. K-means cluster method was used to identify patterns in schedules of different types of common occupations (at home, day work, night work) for each age group which represents the 88% of the United States population. The results obtained detailed information of how occupants spend their time in residential spaces, that allows the creation of occupancy profiles for energy consumption evaluation in the United States buildings.

According to a review in [2], a typical detection system involves multiple environmental sensors, e.g., temperature, carbon dioxide, humidity, infrared, and light, which are distributed throughout a house or space under evaluation. In order to improve the accuracy of occupancy detection techniques, [3] presented an occupancy detection approach employing a two-layer scheme based on data obtained from multiple non-intrusive sensors: temperature, and motion (bottom layer) and application of various Machine Learning (ML) algorithms with human activity modeling to extract occupant information (upper layer). The authors decided to treat the occupancy as a classification problem rather than a regression problem due to the number of occupants in the house (5 people) and the energy consumption that depended on occupancy limit states instead of the number of occupants. These tests were performed in a living lab at Santa Clara University with an area of 65 $m^2$. Four temperature sensors were placed to detect doorknob touches or water usage and a motion sensor (PIR) to detect activity near the entrance door. Also, there were 5 switches, that represented one person per switch, to collect the ground truth of occupant information and a camera for failure detection in the data recollection. The environmental data was gathered over a period of 54 days. Results showed that the accuracy of the ML methods improved from 86% to 95% on average among the studied algorithms when the occupant activity detection layer was implemented.

Application of the traditional Machine Learning (ML) methods have been successful at detecting occupancy patterns for specific datasets. Authors in Ref. [4] proposed an occupancy detection method employing non-intrusive environment data and a Deep-Learning (DL) model. Case study was a university's office. For data collection, they used an environmental sensing board to store information of temperature, humidity, pressure, light level, motion, sound, and carbon dioxide ($CO_2$) and an accurate record of occupancy was established through a ground truth created in a spreadsheet. Five days of data were used to train the DL model. The examination was conducted to assess the

**Digital Object Identifier:** (only for full papers, inserted by LACCEI).
**ISSN, ISBN:** (to be inserted by LACCEI).
**DO NOT REMOVE**

real-time performance of two Convolutional Neural Network (CNN) models deployed on an edge device for the purpose of occupancy detection. The correlation matrix extracted from the analysis revealed that occupancy exhibited the strongest correlation with light level data, while showing the weakest correlation with temperature data. The edge device was then equipped with 1D-CNN and 2D-CNN models, achieving real-time accuracies of 99.72% and 99.76%, respectively.

Ref. [5] state in research relating the occupancy and energy use of a building, the conclusion of most studies are based on a single office or residential building. However, occupancy profiles can have different characteristics depending on the building typology. Therefore, they studied the dynamic relationship between power consumption and WiFi connection counts (as an occupant presence variable) in four buildings located on the National University of Singapore campus. The four building case studies were an office, a lab, a health center and a library where they proposed a new integrated clustering approach to recognize and define the usage patterns across the four buildings typologies. All these buildings had a central district cooling system serving the variable air volume system (HVAC system). The data collection was measured for a year for comparison between the buildings consisting of the following: hourly total building electricity consumption and the hourly number of WiFi connections in the building. Applying linear correlations between electricity consumption and WiFi count data, they found that the metrics had a correlation of 0.93 for the office, 0.88 for the laboratory, 0.67 for the health center, and a 0.74 for the library, except for weekends with a low correlation among the data. They also employed multiple clustering techniques (k-means, DBSCAN, k-shape, HDBSCAN, k-means with PCA, DBSCAN with PCA) and demonstrated that the use of PCA for feature extraction with DBSCAN as cluster method generated distinct clusters with the highest Calinski Harabasz or CH-scores (ratio between cluster variance to within-cluster variance) among the techniques. Furthermore, they concluded that the proposed clustering approach could be used as a preliminary step to better model the variables using methods such as linear regression.

In Ref. [6], the authors state that due to the wide development of Advanced Metering Infrastructure (AMI), high-resolution electricity consumption raw data allows non-intrusive detection of building occupancy by smart meters installed in the customers 'buildings, the electricity consumption data can be registered and transmitted remotely in real-time for occupant detection, therefore it is not necessary to deploy sensors inside the de building (environmental sensors of surveillance cameras). The authors propose a neural network under the name of ABODE-Net which consists of a Deep Learning model of building occupancy detection based on data collection by smart electricity consumption meters. Two databases were used: Electricity Consumption and Occupancy (ECO) and Non-Intrusive Occupancy Monitoring (NIOM). ECO was used to measure five households: one during summer and spring, and the other only during the summer season. For the ECO database an average accuracy of 86.49% and a maximum accuracy of 92.18% were obtained; while for the NIOM database, an average accuracy of 91.76% followed by a maximum accuracy of 93.24% showed that ABODE-Net has a higher capacity than the state of the art reference models in terms of temporal and space utilization for occupancy detection in buildings collected by smart meters, opening up the possibility of implementing this neural network in future studies.

Air conditioning (AC) usage is one of the most popular thermal comforts in residential buildings. Liu, Sun, Mo, conducted in [7] a large-scale measurement study implemented by wireless sensors to monitor indoor environmental parameters such as temperature, relative humidity, and $CO_2$ and $PM_{2.5}$ concentrations between 34 residential buildings of eight cities, from China, distributed in three climate zones: cold, hot summer and cold winter, and hot summer and warm winter. The bedrooms were selected for environmental monitoring and the data were collected for 3 months, from June 2017 to September 2017 including hot and humid climate periods. For AC on/off behavior, they analyzed the AC usage patterns from the occupants (AC power status represented as "0" and "1" for off and on respectively) in the three climates zones and made predictions through models based on logistic regression, artificial neural networks and gradient boosting decision tree algorithms. As results, the AC usage patterns of households in the three climate zones had similarities, the average usage duration and energy consumption of AC per day were 7-10 hours and 2.78 – 3.00 kWh. They concluded that the models could generate an AC operation schedule as an input boundary condition in order to improve simulations of residential buildings energy consumption since the accuracy of prediction from the relative humidity and $CO_2$ concentrations improved the precision of the chosen models.

## II. Materials and methods

### A. Description of Case Study

Case study consists from data collected at two dwellings located near a wooded area in Panama City, Panama, as shown in Fig.1. The studied area in Dwelling 1 consists of a 11.6m$^2$ bedroom on the second floor of a two-story house; it has a south-facing exterior jalousie window, light curtains and a split-type air conditioning system above the window. Kaiterra Sensedge SE-100 indoor air quality (IAQ) monitoring system (Table I) was positioned on a nightstand next to the bed as shown in Fig.2, in order to measure data on indoor air conditions every minute; this data was collected from January 1, 2023 to December 31, 2023, and pre-processing was performed by filling in missing data with the nearest available data and resampling at a 5-minute interval by averaging.

For Dwelling 2, the studied areas consist of the backyard


Fig. 1. Case study location.

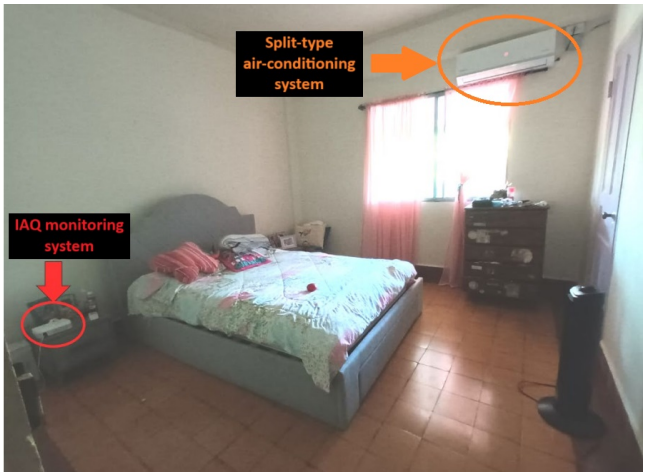


Fig. 2. Positioning of the IAQ monitoring system in the case study bedroom at Dwelling 1.

and the living room on the first floor of a two-story house. Davis Vanguard Pro2 Plus with fan-aspirated radiation shield weather station (Table II) was used to measure indoor and outdoor data; data were collected from May 10, 2023 to August 8, 2023. being pre-processed by filling in missing data with the nearest available data and resampling them at a 15-minute interval by averaging.

K-means clustering techniques were performed on data collected at Dwelling 1 to obtain occupancy and air-conditioning usage profiles. Temperature, relative humidity and carbon dioxide ($CO_2$) were used among the measured variables to perform the clustering algorithm and the number of clusters were decided through Silhouette score analysis. For groundtruth purposes, air-conditioning usage and window opening schedules were implemented as follows: air-conditioning was used from 10:00pm to 5:00am on weekdays and used from 10:00pm to 6:00am on weekends, while window was open from 7:00am to 9:00pm on weekdays and weekends; these schedules are better depicted in Fig.3 and Fig.4.

Additionally, Pearson and Spearman correlation analyses were performed on two databases generated from the data measured at the dwellings: the first database, labeled Case 1, was composed of indoor data collected at Dwelling 1 and

TABLE I
KAITERRA SENSENDGE SE-100 MONITORING SYSTEM SENSOR SPECIFICATIONS [8].

| | | |
|---|---|---|
| $CO_2$ Sensor | Measurement range | 400-2000 ppm |
| | Accuracy | ±3% ±50 ppm |
| | Output resolution | 1 ppm |
| Temperature Sensor | Measurement range | −20°C to 100°C |
| | Accuracy | ±1°C |
| | Output resolution | 0.1°C |
| Humidity Sensor | Measurement range | 0.1% to 99%RH |
| | Accuracy | ±5% |
| | Output resolution | 1% |

TABLE II
DAVIS VANTAGE PRO2 PLUS WEATHER STATION SENSOR SPECIFICATIONS [9].

| | | |
|---|---|---|
| Indoor Temperature Sensor | Measurement range | 0°C to 50°C |
| | Accuracy | ±0.3°C |
| | Output resolution | 0.1°C or 1°C |
| Outdoor Temperature Sensor | Measurement range | −40°C to 65°C |
| | Accuracy | ±0.3°C |
| | Output resolution | 0.1°C or 1°C |
| Dewpoint (calculated) | Measurement range | −76°C to 65°C |
| | Accuracy | ±1°C |
| | Output resolution | 0.1°C or 1°C |
| | Source | World Meteorological Organization (WMO) |
| | Equations used | WMO Equation with respect to saturation of moist air over water |
| | Variables used | Instant Outside Temperature and Current Outside Relative Humidity |
| Wet Bulb (calculated) | Measurement range | −40°C to 65°C |
| | Accuracy | ±1°C |
| | Output resolution | 0.1°C or 1°C |
| | Source | NOAA |
| | Variables used | Current Outside Temperature and Current Outside Relative Humidity |
| Indoor Humidity Sensor | Measurement range | 0.1% to 100%RH |
| | Accuracy | ±2% |
| | Output resolution | 0.1% or 1% |
| Outdoor Humidity Sensor | Measurement range | 0.1% to 100%RH |
| | Accuracy | ±2% |
| | Output resolution | 0.1% or 1% |
| | Drift | < 0.25% per year |
| Solar Radiation Sensor | Measurement range | 0 to 1800 W/m$^2$ |
| | Accuracy | ±5% of full scale[a] |
| | Output resolution | 1 W/m$^2$ |
| | Drift | up to ±2% per year |
| | Cosine response | ±3% for angle of incidence from 0° to 75° |
| | Temperature coefficient | −0.12% per °C[b] |
| Wind Speed Sensor | Measurement range | 0 to 322 km/h |
| | Accuracy | ±3.2 km/h or ±5% |
| | Output resolution | 0.1 km/h or 1 km/h |

[a]Reference: Eppley PSP at 1000 W/m$^2$
[b]Reference temperature = 25°C

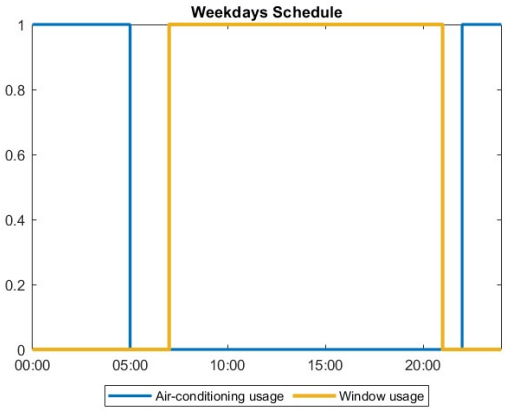


Fig. 3. Weekdays schedules (Air-conditioning usage: from 10:00pm to 5:00am; Window opening: from 7:00am to 9:00pm).

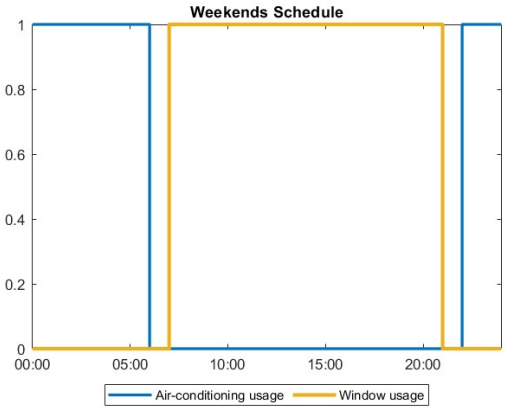


Fig. 4. Weekends schedules (Air-conditioning usage: from 10:00pm to 6:00am; Window opening: from 7:00am to 9:00pm).

outdoor data at Dwelling 2; the second database, labeled Case 2, was composed of indoor and outdoor data measured at Dwelling 2. Both databases were pre-processed by filling in missing data with the nearest available data and resampling them at a 15-minute interval by averaging; temperature, relative humidity, carbon dioxide ($CO_2$), wind speed and solar radiation data were used for these analyses.

### B. Correlation coefficient analysis

Correlation coefficient analysis is a statistical analysis that measures the degree of interrelation between variables; it can take values between $-1$ and $1$. A correlation of $1$ indicates that the variables are perfectly positively correlated, a correlation of $-1$ means that the variables are perfectly negatively correlated and a correlation of $0$ means that the variables are not correlated [10]–[12].

In this paper we use Pearson's correlation coefficients and Spearman's rank correlation coefficients analysis to study the relationships between the measured variables. Pearson's correlation coefficient analysis describes the linear relationships that exist between the variables [13], [14], while Spearman's rank correlation coefficient analysis is used to determine the degree of monotonic relationship between the variables [15], [16].

### C. k-Means clustering algorithm

Clustering algorithms are a set of techniques used to group unlabeled databases by their similarities, so that data from the same group are as similar as possible and data from different groups are as different as possible. k-Means is one of the most widely used and popular clustering techniques, it groups data by establishing centroids and assigning each data point to the centroid of the nearest cluster, the number of centroids is decided by the user through the number of clusters (parameter k) [17]–[19].

The k-means clustering algorithm is as follows: first, the algorithm randomly locates k centroids in the database and then alternates between assigning the data points to the nearest cluster and locating new centroids using the newly created clusters. This cycle is repeated until the distances in each cluster are minimized and data points are not reassigned to new clusters [17], [18].

### D. Silhouette score analysis

Selecting the best parameters to implement a clustering algorithm can be a difficult task to perform. Validation techniques such as Calinski-Harabasz values or Silhouette score values can be used for this purpose; in this work we use Silhouette score values to decide the best number of clusters for each type of data.

Silhouette score values are an indicator of how similar a data point is to others in the same cluster compared to the other clusters. These values are denoted by (1), where $a_i$ is the average distance from the *i*-th point to the other points in the same cluster and $b_i$ is the minimum average distance from the *i*-th point to points in a different cluster, minimized over the clusters [20].

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{1}$$

For the Silhouette score analysis, an average of all the Silhouette values of the cluster is obtained, this average ranges from -1 to 1 and is used to indicate the quality of the cluster: the higher the average, the better the clustering parameters used and the better the solution obtained.

## III. Results and Discussion

### A. Silhouette score analysis

Silhouette score analysis was implemented for the decision of the number of clusters for each variable. Fig.5 and Fig.6 show that using two clusters presents a better score for carbon

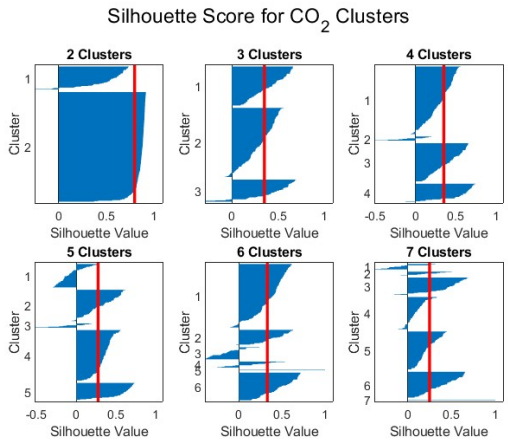


Fig. 5. Silhouette score for carbon dioxide ($CO_2$) clusters.

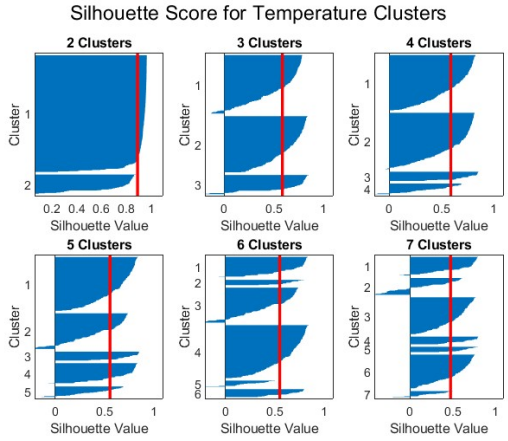


Fig. 6. Silhouette score for temperature clusters.

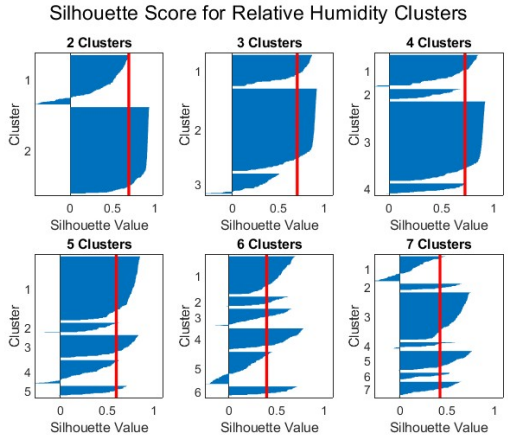


Fig. 7. Silhouette score for relative humidity clusters.

dioxide and temperature respectively. For relative humidity, the best scores were obtained using two, three, and four clusters. However, we decided to employ two clusters for this variable to our results, since it groups the data in a more distinctive

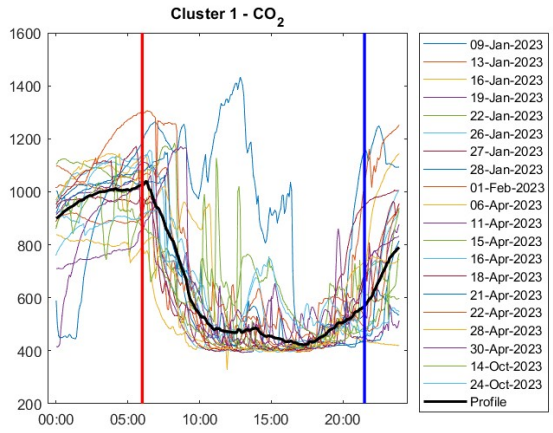


Fig. 8. Indoor $CO_2$ cluster 1 (Red vertical line: turning off the air conditioning system and opening the window; Blue vertical line: turning on the air conditioning system and closing the window).

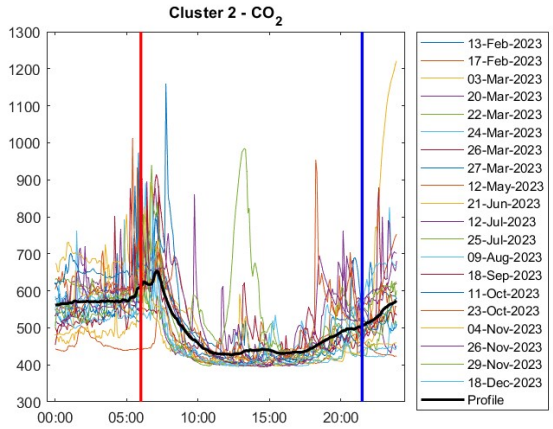


Fig. 9. Indoor $CO_2$ cluster 2 (Red vertical line: turning off the air conditioning system and opening the window; Blue vertical line: turning on the air conditioning system and closing the window).

manner, with clusters not having the same behavior between them and a better adaptation to the results we were looking for.

### *B. k-Means clustering technique*

For $CO_2$ data, the resulting clusters are as follows: Cluster 1, shown in Fig.8, groups days with high $CO_2$ concentrations during the night and low concentrations during the day; Cluster 2, shown in Fig.9, groups days with medium $CO_2$ concentrations during the night and low concentrations during the day. Analyzing the resulting $CO_2$ clusters, no multiple occupancy profiles can be inferred from them, as both clusters describe the same occupancy behavior: a nighttime occupant presence and a daytime occupant absence; obtaining similar results as authors in [4], both of them suggest that $CO_2$ can be used to estimate the occupant behavior. Air conditioning usage profiles are detailed as follows: Cluster 1 describes days of nighttime air conditioning use and daytime open window

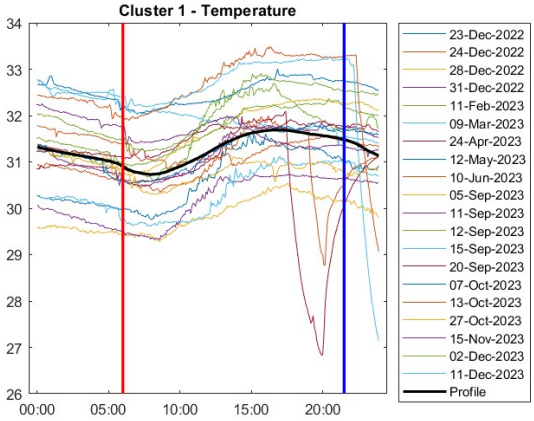


Fig. 10. Indoor temperature cluster 1 (Red vertical line: turning off the air conditioning system and opening the window; Blue vertical line: turning on the air conditioning system and closing the window).

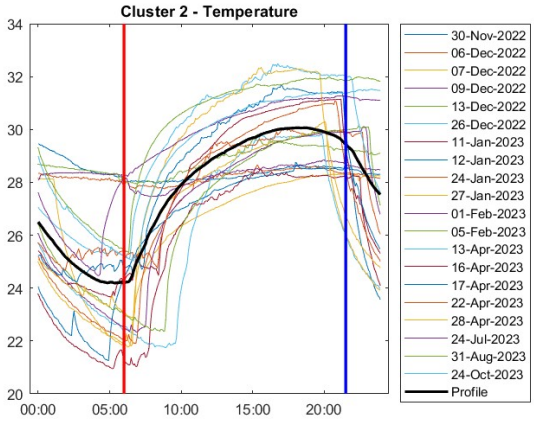


Fig. 11. Indoor temperature cluster 2 (Red vertical line: turning off the air conditioning system and opening the window; Blue vertical line: turning on the air conditioning system and closing the window).

use, while Cluster 2 describes days of daytime and nighttime open window use.

The temperature and relative humidity groups are less informative about the presence of occupants in the bedroom, as air conditioning uses affected these variables more than occupancy changes; this results in only multiple air conditioning usage profiles being extracted from the clusters for these variables. Clusters from temperature data can be described as follows: Cluster 1, depicted in Fig.10, groups days with consistently high temperatures during the day and night; this cluster describes days when no air conditioning was used. Cluster 2, depicted in Fig.11, groups days with low temperatures at night and high temperatures during the day; this cluster describes days when air conditioning was used during the night.

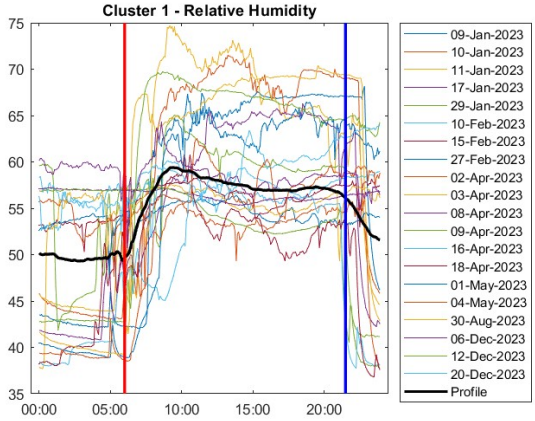


Fig. 12. Indoor relative humidity cluster 1 (Red vertical line: turning off the air conditioning system and opening the window; Blue vertical line: turning on the air conditioning system and closing the window).

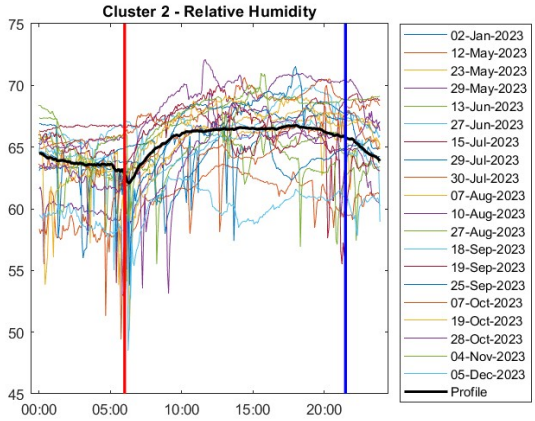


Fig. 13. Indoor relative humidity cluster 2 (Red vertical line: turning off the air conditioning system and opening the window; Blue vertical line: turning on the air conditioning system and closing the window).

As for relative humidity data, the clusters showed the following behaviors: Cluster 1, presented in Fig.12, groups days with low relative humidity at night and high relative humidity during the day; this cluster describes days when air conditioning was used during the night. Cluster 2, presented in Fig.13, groups days with constant high relative humidity during the day and night, this cluster describes days when air conditioning was not used.

The implementation of $CO_2$ and relative humidity data to obtain air conditioning usage profiles is supported by [7]; the results obtained in this study showed that using these features considerably improved the accuracy of the models generated by the different algorithms.

### C. Correlation Coefficient Analysis

For Case 1, which is the case study, Fig.14 and Fig.15 illustrate the Pearson and Spearman correlation analyses,

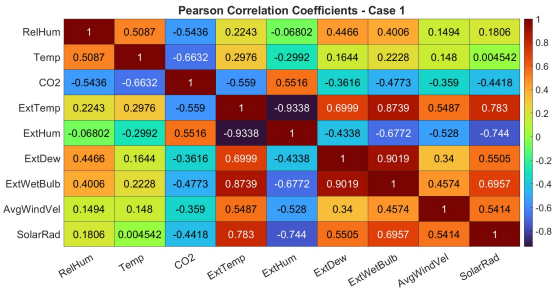


Fig. 14. Pearson Correlation Coefficients for Case 1.

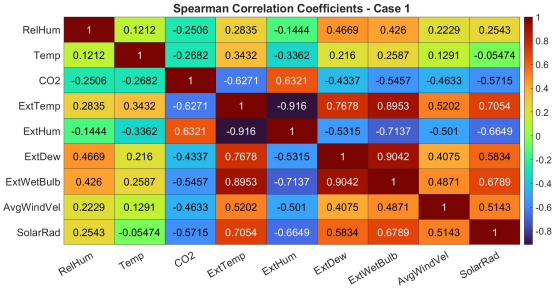


Fig. 15. Spearman Correlation Coefficients for Case 1.

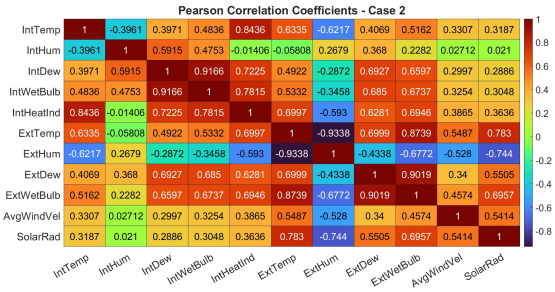


Fig. 16. Pearson Correlation Coefficients for Case 2.

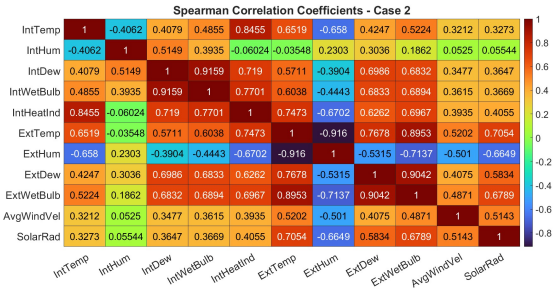


Fig. 17. Spearman Correlation Coefficients for Case 2.

respectively. They show that only $CO_2$ is highly correlated with the outdoor variables, which is to be expected since outdoor conditions can influence the behavior of the occupants and their carbon dioxide production. When evaluating the correlation coefficients between the indoor variables, it is observed that the variables have low Spearman coefficients and high Pearson coefficients, which mean that the relationship between variables shows a higher linear behavior, but they are not predominantly increasing or decreasing.

As for Case 2, Fig.16 and Fig.17 represent the heat map of Pearson and Spearman correlation analyses, respectively. They showed that almost all outdoor variables are strongly correlated with indoor temperatures but not with indoor relative humidity, which means that changes in outdoor conditions affect indoor temperatures more than indoor humidity. Analyzing the correlation coefficients between the indoor variables, it is observed that indoor temperature is not strongly correlated with indoor relative humidity, but indoor dew and wet bulb temperatures are strongly correlated with indoor humidity; these strong correlations are to be expected since both temperatures are based on ambient humidity.

## IV. Conclusion

This research employed the k-means clustering technique to the following features: temperature, relative humidity, and carbon dioxide for data collection in order to obtain profiles of occupancy and use of air conditioning in the room.

The observations, each of them representing one day, were grouped into two clusters, which represent air conditioning use profiles: both show an open window usage during daytime; however, during nighttime, one profile stands for air conditioning usage while the other maintains open window usage. Different occupancy profiles could not be generated from the study features, since there was no difference in the occupancy behavior from the clustering results; nevertheless, analysis of the carbon dioxide data was able to represent daily occupant behavior, having an occupant presence during nighttime and an occupant absence during daytime.

Future work includes the collection of room environmental data from several points to study the benefits of collecting data at one or different points in through the area, or implementing one of the algorithms mentioned in the Introduction to perform a comparison between algorithms.

## Acknowledgment

The authors extend their gratitude to the Research Group Energética y Confort en Edificaciones Bioclimáticas of the Faculty of Mechanical Engineering within the Universidad Tecnológica de Panamá. This publication is part of the project FID22-086, which has received funding from Secretaría Nacional de Ciencia, Tecnología e Innovación (SENACYT), together with the Sistema de Nacional de Investigación (SNI).

## References

[1] D. Mitra, Y. Chu, and K. Cetin, "Cluster analysis of occupancy schedules in residential buildings in the United States," vol. 236, p. 110791. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S037877882100075X

[2] L. Rueda, K. Agbossou, A. Cardenas, N. Henao, and S. Kelouwani, "A comprehensive review of approaches to building occupancy detection," vol. 180, p. 106966. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360132320303255

[3] C. Wang, J. Jiang, T. Roth, C. Nguyen, Y. Liu, and H. Lee, "Integrated sensor data processing for occupancy detection in residential buildings," vol. 237, p. 110810. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378778821000943

[4] A. N. Sayed, F. Bensaali, Y. Himeur, and M. Houchati, "Edge-Based Real-Time Occupancy Detection System through a Non-Intrusive Sensing System," vol. 16, no. 5, p. 2388. [Online]. Available: https://www.mdpi.com/1996-1073/16/5/2388

[5] S. Zhan and A. Chong, "Building occupancy and energy consumption: Case studies across building types," vol. 2, no. 2, pp. 167–174. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2666123320300829

[6] Z. Luo, R. Qi, Q. Li, J. Zheng, and S. Shao, "ABODE-Net: An Attention-based Deep Learning Model for Non-intrusive Building Occupancy Detection Using Smart Meter Data," vol. 13828 LNCS, pp. 152–164.

[7] H. Liu, H. Sun, H. Mo, and J. Liu, "Analysis and modeling of air conditioner usage behavior in residential buildings using monitoring data during hot and humid season," vol. 250, p. 111297. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378778821005818

[8] Sensedge — Premium Indoor Air Quality Monitor with Display — Kaiterra. [Online]. Available: https://www.kaiterra.com/sensedge

[9] Wireless Vantage Pro2 Plus with 24-Hour Fan Aspirated Radiation Shield and WeatherLink Console - SKU 6263, 6263M. Davis Instruments. [Online]. Available: https://www.davisinstruments.com/products/wireless-vantage-pro2-plus-with-24-hr-fan-aspirated-radiation-shield-and-weatherlink-console

[10] S. K. Haldar, "Chapter 9 - Statistical and Geostatistical Applications in Geology," in *Mineral Exploration (Second Edition)*, S. K. Haldar, Ed. Elsevier, pp. 167–194. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128140222000095

[11] S. Mishra and A. Datta-Gupta, "Chapter 2 - Exploratory Data Analysis," in *Applied Statistical Modeling and Data Analytics*, S. Mishra and A. Datta-Gupta, Eds. Elsevier, pp. 15–29. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B978012803279400002X

[12] N. Sharma and R. Deo, "Chapter 14 - Wind speed forecasting in Nepal using self-organizing map-based online sequential extreme learning machine," in *Predictive Modelling for Energy Management and Power Systems Engineering*, R. Deo, P. Samui, and S. S. Roy, Eds. Elsevier, pp. 437–484. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128177723000148

[13] V. A. Profillidis and G. N. Botzoris, "Chapter 5 - Statistical Methods for Transport Demand Modeling," in *Modeling of Transport Demand*, V. A. Profillidis and G. N. Botzoris, Eds. Elsevier, pp. 163–224. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128115138000054

[14] D. Chen and C. J. Anderson, "Categorical data analysis," in *International Encyclopedia of Education (Fourth Edition)*, R. J. Tierney, F. Rizvi, and K. Ercikan, Eds. Elsevier, pp. 575–582. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128186305100703

[15] H. Bon-Gang, "Chapter 3 - Methodology," in *Performance and Improvement of Green Construction Projects*, H. Bon-Gang, Ed. Butterworth-Heinemann, pp. 15–22. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B978012815483000003X

[16] K. K. Al-jabery, T. Obafemi-Ajayi, G. R. Olbricht, and D. C. Wunsch II, "2 - Data preprocessing," in *Computational Learning Approaches to Data Analytics in Biomedical Applications*, K. K. Al-jabery, T. Obafemi-Ajayi, G. R. Olbricht, and D. C. Wunsch II, Eds. Academic Press, pp. 7–27. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128144824000024

[17] CS221. [Online]. Available: https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

[18] K-Means Cluster Analysis — Columbia Public Health. Columbia University Mailman School of Public Health. [Online]. Available: https://www.publichealth.columbia.edu/research/population-health-methods/k-means-cluster-analysis

[19] What is K Means? NVIDIA Data Science Glossary. [Online]. Available: https://www.nvidia.com/en-us/glossary/k-means/

[20] Silhouette plot - MATLAB silhouette - MathWorks América Latina. [Online]. Available: https://la.mathworks.com/help/stats/silhouette.html?s_tid=doc_ta