Environmental Science: Atmospheres

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Environ. Sci.: Atmos.*, 2025, **5**, 1144

An unsupervised machine learning approach for indoor air pollution analysis

Bárbara A. Macías-Hernández, †* Edgar Tello-Leal, †* Jailene Marlen Jaramillo-Perez and René Ventura-Houle

Exposure to indoor air pollutants is one of the most significant environmental and health risks people face, especially since they spend most of their time indoors. Therefore, evaluating indoor air pollution levels and comfort parameters is essential for achieving sustainable indoor air quality (IAQ). The main objective of this study was to identify patterns of indoor air pollution in two buildings with different characteristics located on a university campus in northeastern Mexico. We measured the concentration of particulate matter in fractions of 1.0 μm (PM_1), 2.5 μm ($PM_{2.5}$), and 10 μm (PM_{10}), as well as carbon dioxide (CO_2), carbon monoxide (CO), and ozone (O_3), along with the temperature and relative humidity in each microenvironment during the working hours of spring, summer, and autumn. Next, unsupervised machine learning was employed to identify behavioral patterns of air pollutants within the microenvironments. The *K*-means clustering algorithm was used to identify homogeneous microenvironments within the study area. We performed three clustering analyses per building: (1) considering all the variables in the dataset, (2)selecting the significant variables through principal component analysis (PCA), and (3) examining two time ranges within the working day. The robustness of the proposed approach was evaluated through a comparative analysis of the *K*-means, DBScan, and hierarchical algorithms, assessing their performance using the Davies–Bouldin index and Silhouette score metrics. Furthermore, the stability of the clusters over time intervals was assessed using the adjusted Rand index. Cluster analysis enabled us to identify microenvironments with maximum similarity and those that change groups, as their behavior depends on the time range. Consequently, grouping microenvironments into homogeneous IAQ classes is effective in accurately identifying spaces based on patterns related to their contamination levels and guiding actions to reduce pollution levels by zone or building.

Received 25th April 2025
Accepted 6th September 2025

DOI: 10.1039/d5ea00051c

rsc.li/esatmospheres

Environmental significance

Indoor air pollution represents a serious problem due to its direct impact on individuals. In developing countries, its effects are more pronounced because of limited resources for monitoring and subsequent delays in taking action to reduce air pollutant concentration levels. Prolonged exposure to air pollutants is suspected of causing serious health effects. In the case of indoor air pollution, the risk increases, as most people worldwide spend the majority of their time indoors. Clustering analysis enabled us to identify the microenvironments with the greatest similarity (homogeneous classes) and those that transition between groups due to their behavioural patterns, which depend on air pollution concentration within a specific time range.

1 Introduction

Poor indoor air quality (IAQ) can impact a person's health, comfort, cognitive performance, and work capacity, resulting in productivity losses.[1,2] Therefore, IAQ analysis is essential in achieving a sustainable and healthy urban environment.[3] Recently, several studies on air quality have been presented in schools, office buildings, dwellings, hospitals, and other workspaces.[4–7] Allowing identified indoor air pollution sources, leading pollution problems, possible causes, levels of air pollution and their assessment, and recommendations to improve IAQ. Air pollutants commonly identified inside buildings are ozone (O_3), particulate matter ($PM_{2.5}$ and PM_{10}), carbon monoxide (CO), carbon dioxide (CO_2), volatile organic compounds (VOCs), and bioaerosols.[8–10]

Factors related to temperature, humidity, inadequate ventilation, mold caused by humidity, lighting, and exposure to chemical substances are considered in determining this quality.[11–13] In this sense, inadequate ventilation with low air exchange rates can influence elevated CO_2 concentrations in offices and increase the concentration of other indoor air pollutants.[14–17] Temperature and relative humidity levels are

Faculty of Engineering and Science, Autonomous University of Tamaulipas, Victoria 87000, Mexico. E-mail: etello@docentes.uat.edu.mx

† These authors contributed equally to this work.

strongly associated with comfort and are the main factors influencing occupant productivity.[18] High levels of humidity and low or high temperatures inside an office affect the cooling capacity of the human body, causing health problems,[19] such as irritation of the eyes, nose, and throat, headaches, dizziness, and fatigue.[20,21] These symptoms also relate to the highest CO_2 levels in the indoor environment.[19,22,23] Long-term exposure to $PM_{2.5}$ and PM_{10} negatively affects the respiratory, cardiovascular, and nervous systems and, in recent studies, has been associated with high mortality rates.[24–26] Additionally, asthma, coughing, wheezing, shortness of breath, sinus congestion, sneezing, nasal congestion, and sinusitis have been associated with long and short-term exposure to $PM_{2.5}$, PM_{10}, CO, O_3, and VOC.[25–28] Although people usually react differently to indoor air pollutants, prolonged exposure to high concentrations can lead to specific health conditions.[29]

The sources that generate poor air quality inside a building vary according to the activities carried out, the consumption of products, and the building's location,[30–32] making it a complex situation to analyze.[9] For example, O_3 is commonly emitted by electronic office equipment such as laser printers and photocopiers.[33,34] Hence, indoor pollutants can be emitted from multiple sources in low concentrations, resulting in mixed air conditions and a greater health risk. CO, benzene, sulfur dioxide (SO_2), O_3, nitrogen oxides (NO_x), $PM_{2.5}$, and PM_{10} can be mentioned in the pollutants from external sources. Internal source pollutants have been identified, including CO_2, bioefluents, $PM_{2.5}$, and PM_{10}, as well as construction-related contaminants such as microbials and VOCs.

Therefore, recognizing microenvironments that share similar characteristics of air pollutant concentration and comfort parameters is a current requirement. This enables the identification of shared pollution sources within a building and the development of solutions, policies, or actions that reduce concentrations. Several research studies have confirmed the robustness of clustering approaches (mainly the *K*-means algorithm) in air quality management,[35,36] discovering daily patterns of IAQ,[37] assessing fluctuations in indoor thermal conditions,[38] an indoor fine and ultrafine particle clustering method,[39] identification of IAQ events from gas sensor data,[40] and indoor/outdoor air pollution analysis.[41,42] Clustering algorithms are machine learning techniques classified within unsupervised learning approaches, where the algorithm implemented has no prior knowledge of the class to which the instances belong and attempts to extract patterns or behaviors that allow for the inference of some relationship (similarity) between the cases in the dataset. Despite the drawbacks widely discussed in the literature,[43,44] the *K*-means clustering algorithm is one of the most reliable and widely used approaches in unsupervised learning.[45,46] Cluster analysis aims to identify objects with high similarity by creating partitions between the instances based on the data (values) of the characteristics that make up the dataset; that is, data clustering aims to divide unlabeled data into groups based on the distance measure that makes it possible to identify the similarity between them.[47]

The purpose of this research is to identify and visualize the groups of indoor microenvironments that share similar patterns from the concentration levels of pollutants PM_1, $PM_{2.5}$, PM_{10}, CO, CO_2, O_3, and air comfort parameters using the unsupervised learning algorithm *K*-means within a university campus in northeastern Mexico. Furthermore, the results of IAQ monitoring in microenvironments, including offices and workspaces, distributed across two buildings with different construction characteristics, are presented. The data was collected for eight continuous hours (per office) during the working day in the spring, summer, and autumn seasons of 2023.

2 Methods

2.1 Study population

In our experiment, we monitored the indoor air quality (IAQ) in two buildings of the Faculty of Engineering and Sciences in Victoria City, Tamaulipas, Mexico. These buildings were selected because they are the locations that concentrate the most significant number of administrative staff and professors of the institution. Building A houses the research and postgraduate division offices on the fourth level of a seven-story building with glass walls, built in 2012. It has 11 ventilation and air conditioning systems installed on the fourth floor, each equipped with filters that meet the ASHRAE MERV-8 standard. These filters can capture up to 90% of particles measuring 3 to 10 micrometers and are designed to trap common indoor pollutants, such as mold spores, dust, and pet dander. Professors primarily use the offices in building A, while building B corresponds to the central administration offices of the faculty. This building was constructed on one floor in 1967. This building has four ventilation and air conditioning systems installed with MERV-8 filters and five units with filters that only meet a minimum standard. The administrative staff carries out all school management activities in this building. The study examined 26 microenvironments, including 13 for buildings, to monitor indoor air quality (IAQ) conditions. In building A, data were collected in nine offices and four workspaces, with 18 participants involved in the study. In building B, nine offices and four workspaces were monitored, with 23 people cooperating.

The buildings considered in our study are smoke-free workspaces. Each microenvironment typically has an average of three devices: a personal computer, a laser or inkjet printer, a photocopier, an IP phone, a paper shredder, surveillance and access point devices, and interconnection equipment to the data network. Building A does not have access to natural ventilation, and its exterior walls are made of glass, making it a hermetic building. A central air conditioning system provides ventilation, regulating the interior temperature for the entire building. On the other hand, in building B, ventilation is achieved through a central air conditioning system, allowing personnel to regulate the interior temperature. It is possible to have natural ventilation in the autumn and early wintertime by opening the windows.

2.2 Sample collection

The indoor air pollutant gasses data were measured continuously using sensors installed in a Libelium Smart Environment

Pro air quality monitoring station,[48] and the concentration of particulate matter (PM_1, $PM_{2.5}$, PM_{10}) were calculated with a Plantower PMS7003 sensor.[49] The meteorological parameters were measured with a BOSCH BME280 sensor.[50] The particulate matter, CO, and O_3 sensors were calibrated through a collocation process with a regulatory-grade reference instrument or equivalent monitor (FRM/FEM) under real-world conditions during an evaluation period specified in the US EPA's methodology for low-cost sensor calibration.[51,52] The O_3 sensor achieved a coefficient of determination (R^2) of 0.90; the particulate matter sensor achieved an R^2 of 0.96 for PM_1, 0.90 for $PM_{2.5}$, and 0.88 for PM_{10}; and the CO sensor achieved an R^2 of 0.91. The calibration process for the PM_1, $PM_{2.5}$, PM_{10}, O_3, and CO sensors was similar to that described in our previous work.[53] The CO_2 sensor was calibrated in a laboratory using a vacuum chamber with a controlled internal temperature of 20 °C and a relative humidity of 55% (±5%), resulting in an R^2 of 0.91. The measurements were taken continuously for eight hours indoors at 3 minutes intervals during the workday (9:00 a.m. to 5:00 p.m.), collecting data for each microenvironment in each of the four seasons of the year. The air quality monitoring station was mounted on a height-adjustable tripod (to ensure the desired height) and placed approximately 1.3 to 1.5 meters high, matching the breathing zone of a sitting person. The sampling occurred from April to May (spring), July to August (summer), and October to November 2023 (autumn).

2.3 Data analysis

Data was processed using the statistical analysis software R-Studio IDE 4.3.3 through the Posit Cloud. Since the data exhibited a non-normal distribution, the Spearman rank correlation analysis was performed to discover the degree of association between the study variables. The following tasks were applied to implement the unsupervised learning approach using the *K*-means clustering technique. The datasets were normalized using the BB-misc:normalize library. The distance or proximity between two objects was calculated using the Euclidean distance, which defines their similarity. Next, a principal component analysis (PCA) is applied to select the significant variables contributing to each principal component and explain the maximum variance of all the variables, helping to reduce high dimensionality. The PCA was implemented using the FactoMineR library version 2.11. In each experiment, the optimum cluster value (K) was calculated using the Elbow method using the fviz_nbclust function of the factoextra library version 1.0.7. Finally, the clustering algorithm was implemented using the *K*-means function of the Stats 4.3.3 library, and the visualization of the generated clusters was done using the fviz_cluster function of the factoextra library. Evaluating the quality of each generated model using the silhouette score metric, using the cluster library version 2.1.6. A comparative analysis was also conducted between the *K*-means, DBSCAN, and hierarchical algorithms to assess the robustness of the proposed method, using the Davies–Bouldin index and Silhouette score metrics for performance evaluation. Additionally, the stability of the clusters over different time intervals was examined with the adjusted Rand index. Finally, the loading changes over intraday time ranges of the study variables were analyzed using PCA.

3 Results

3.1 Correlation analysis

The Spearman correlation analysis was used to decipher the relationships between variables of indoor air pollutants and meteorological variables. Fig. 1–3 show correlation matrices generated with the Spearman method corresponding to the spring, summer, and autumn seasons, respectively. The relationship between CO and relative humidity (RH) is a moderate negative to moderate positive correlation ($-0.41 \geq r_s \geq 0.59$), with a statistically significance p-value < 0.05, in the spring and summer for building A (see Fig. 1 and 2). The shift from a negative to a positive correlation suggests that underlying factors influence the relationship between CO and RH, and these factors change over time or in response to environmental conditions. Possible factors include ventilation and air conditioning systems, CO sources, environmental dynamics, and occupant behavior during daily activities.

Likewise, an association with moderate strength was identified between CO and O_3 in the summer period ($r_s = 0.41$, p-value < 0.05). This correlation suggests that both pollutants exhibit similar dynamics within the building, driven by factors such as outdoor air filtration and the operation of ventilation systems during the summer, which result in a simultaneous increase in their concentrations. In the analysis carried out for the microenvironments of building A, we observe that the O_3 and RH exhibit a consistently moderate negative to strong positive relationship in the spring of $r_s = -0.48$, summer of $r_s = 0.66$, and autumn of $r_s = 0.75$ periods (see Fig. 3). The strongly positive correlations identified in autumn and summer result

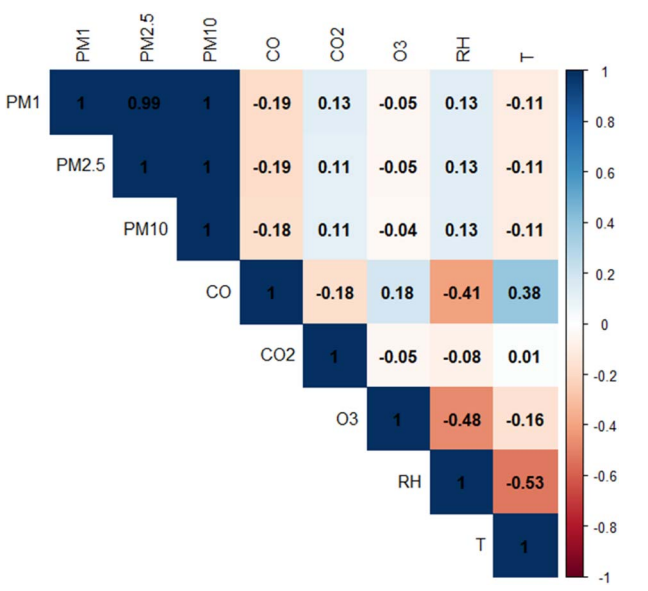

Fig. 1 Visualization of the Spearman coefficient correlation matrix for the spring season using the building A dataset.

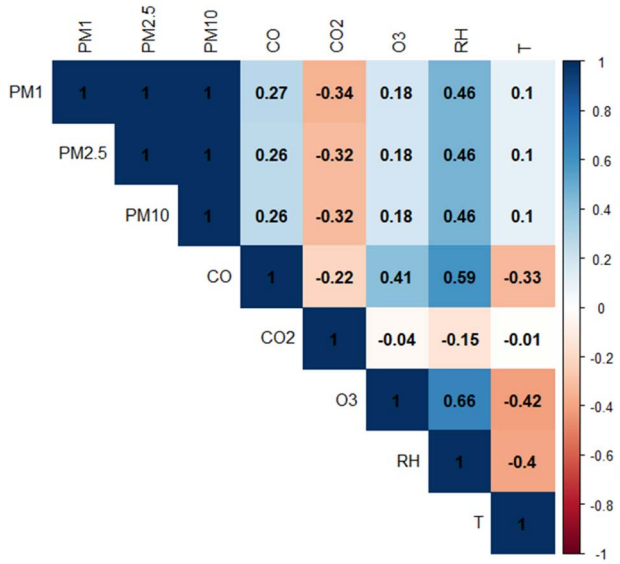

Fig. 2 Visualization of the coefficient correlation matrix for the summer season for building A.

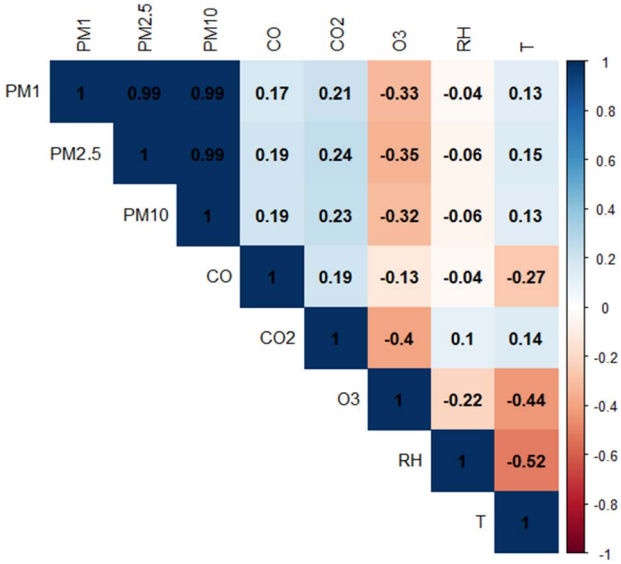

Fig. 3 Visualization of the Spearman coefficient correlation matrix for the spring season using the building B dataset.

from the hot, sunny days typical of these seasons. These times usually have the highest outdoor O_3 levels, and outdoor relative humidity is also very high. As a result, the HVAC system must draw in outside air to ventilate the building, actively introducing air that contains both O_3 and humidity. During autumn, the magnitude of the correlation coefficient between temperature (T) and O_3 reaches its highest value ($r_s = -0.81$, p-value > 0.05) in the correlation matrix, confirming a very strong correlation (see Fig. 3). This relationship is also observed in the summer, but with a moderate negative strength. The variables for particulate matter in its fractions (PM_1, $PM_{2.5}$, PM_{10}) present a moderate association with the RH, both in summer ($r_s = 0.46$) and in autumn ($r_s = 0.54$). Finally, Fig. 1–3 show that the RH and T variables establish a relationship with a coefficient between −0.40 and −0.59 in the three seasons analyzed.

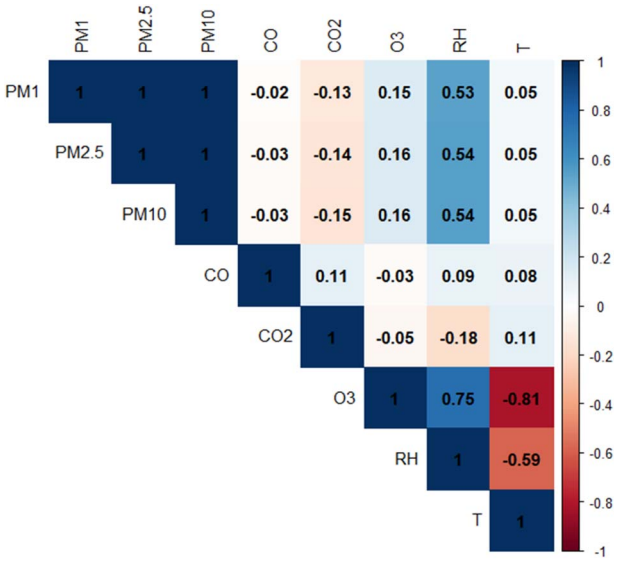

Fig. 4 Visualization of the Spearman coefficient correlation matrix for the autumn season for building A dataset.

Fig. 4 shows a correlation matrix for the spring season of the building B dataset. In this figure, three moderate negative associations are observed between the variables of CO_2 and O_3 ($r_s = -0.40$), T and O_3 ($r_s = -0.44$), and T and RH ($r_s = -0.52$). More-over, moderate negative relationships were identified between the O_3 and T variables of $r_s = -0.42$ and $r_s = -0.58$ in summer and autumn (see Fig. 5 and 6), respectively. Furthermore, we found that CO has a moderate negative correlation with the T variable in the summer and autumn periods, and

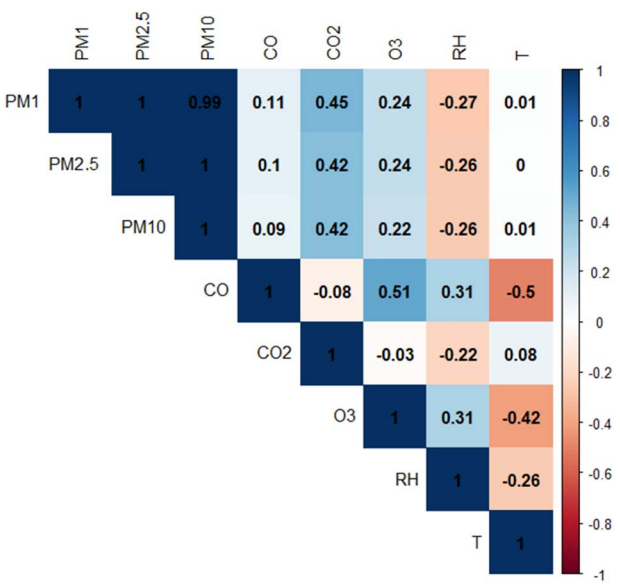

Fig. 5 Visualization of the coefficient correlation matrix for the summer season for building B.

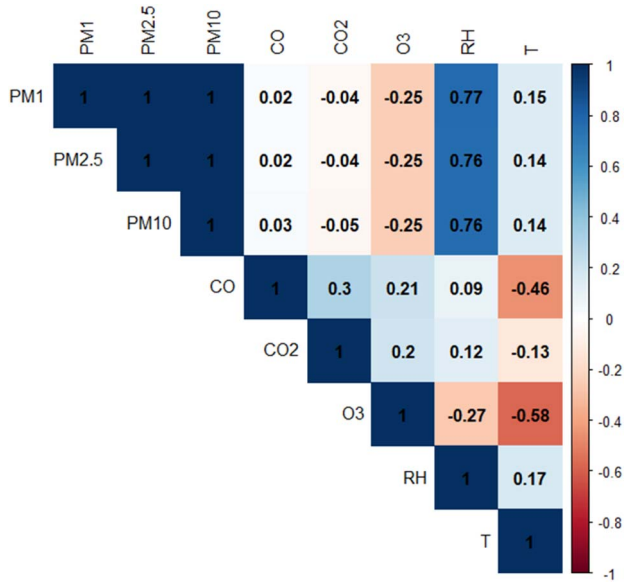

Fig. 6 Visualization of the Spearman coefficient correlation matrix for the autumn season for building B dataset.

a moderate positive association was observed between the CO and O_3 variables with a coefficient of $r_s = 0.51$. Finally, the PM_1, $PM_{2.5}$, and PM_{10} variables are associated with moderate strength with the CO_2 variable, with $r_s = 0.45$, $r_s = 0.42$, and $r_s = 0.42$, respectively. The positive and moderate correlation between CO_2 and particulate matter fractions suggests that these variables move in the same direction. This is highly relevant in the context of a building. One possible explanation is that poor ventilation or high occupancy in offices leads to a simultaneous increase in CO_2 levels (exhaled by occupants) and the resuspension of particles generated by human activities, along with filtration from outside.

On the other hand, in the autumn period, the highest coefficient was found in the correlation matrix of building B, with a strong positive association between the PM_1 and RH variables of $r_s = 0.77$ (see Fig. 6). Further, the $PM_{2.5}$ and PM_{10} variables are related to RH but with a value of $r_s = 0.76$ with p-value < 0.05. The strong positive correlation suggests that relative humidity is a key factor in indoor particle concentrations during the fall, likely due to the hygroscopic properties of the particles or their interaction with ambient conditions typical of this season.

3.2 Clustering analysis

In the first experiment with the data from building A, using all the variables and the data from the three seasons of the year considered in the study, 3 clusters were obtained, consisting of 20 elements in cluster 1, 13 objects in cluster 2, and 6 elements in cluster 3 (see Fig. 7). Objects are identified by office number and the season of the year (*e.g.*, 8–2 for office 8 in the summer). In cluster 1, it was observed that instances presenting Office 12 in the three seasons exhibit high similarity, forming part of the same group. Furthermore, the cases of offices 3, 6, 8, 9, and 11 in the summer and autumn seasons exhibit similar

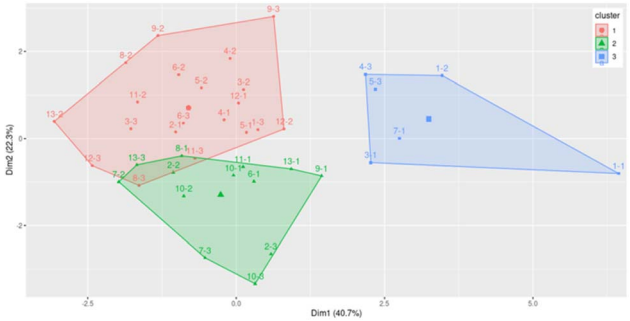

Fig. 7 Clusters generated from the data of building A, considering the 8 study variables.

characteristics, and a minimum distance is observed between the observations of offices 4 and 5 in spring and summer. In Fig. 7, it can be observed that objects 8–3 and 11–3 of cluster 1 and objects 8–1, 2–2, 7–2, and 13–3 of cluster 2 are very close, with a high similarity, despite belonging to different clusters. The elements grouped in cluster 3 are characterized by sharing high values in the con-centration levels of PM_1, $PM_{2.5}$, PM_{10}, and O_3, as well as a high percentage of relative humidity.

In a second experiment, principal component analysis (PCA) was implemented to capture the relevant and highly correlated variables, as well as to eliminate noisy variables and reduce over-fitting, thereby improving the performance of the algorithm when forming clusters with the microenvironment data. Fig. 8 displays the scree plot, which illustrates the significance of each principal component in building A dataset. This Fig. 8 displays the percentage of explained variances by each component. The first component almost explains 41% of the total variance, implying that the first principal component can represent nearly two-quarters of the data from the eight variables. In this case, four components are required to explain 85.73% of the total information in the data with the variables CO, $PM_{2.5}$, PM_{10}, PM_1, and CO_2. If five principal components are considered, 94.54% of the accumulated variance is reached, and the variables with the greatest contribute are CO_2, O_3, CO, $PM_{2.5}$, PM_{10}, PM_1.

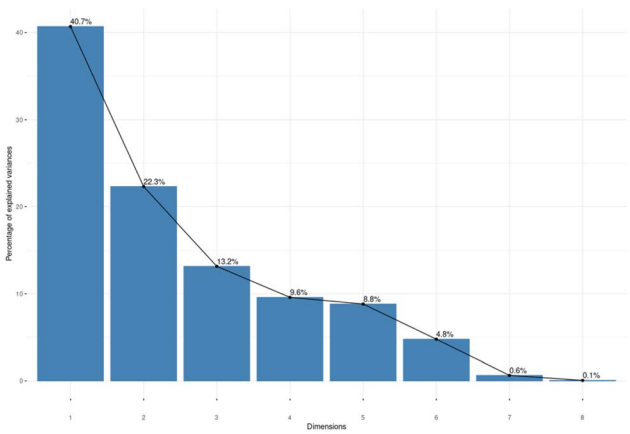

Fig. 8 Scree plot of the components in the data analysis of building A.

Then, a sub-dataset containing the PM_1, $PM_{2.5}$, PM_{10}, CO, and CO_2 variables was built to analyze them using the clustering algorithm. Fig. 9 shows the groups created with a building A dataset. In this figure, the overlap between elements of clusters 1 and 2 is no longer observed (compared to Fig. 7). Objects 8–1, 11–1, 2–2, and 7–2 changed their membership to cluster 1, leaving cluster 2, as mentioned in the analysis of Fig. 7. Likewise, objects 5–1, 12–1, 12–2, and 1–3 were added to cluster 2. Cluster 3 maintains the same elements, incorporating a clear pattern of behavior, and includes an object (9–1) from cluster 2. The *K*-means algorithm detected patterns more clearly in the dataset, particularly when working with low-dimensional data. Even though the dataset used does not contain many variables (high dimensionality), the PCA function helps discard noisy and non-significant variables to discover behavioral patterns.

The third experiment analyzes the data by time range: 8 a.m. to 12 p.m. will be referred to as time-1, and 1 p.m. to 5 p.m. will be referred to as time-2. Fig. 10 and 11 show the objects that comprise the clusters in time-1 and time-2, respectively. When comparing the elements that make up the clusters in the two-time ranges, it was identified that 62% of the objects maintain their membership in the same group. In the cluster analysis for the time-1 range, four components were considered that reached a percentage of variance explanation of 84.28%, and the contribution of the variables has the following order: $PM_{2.5}$, PM_{10}, PM_1, CO and CO_2. In the case of the PCA analysis used in the clustering of time-2, it is made up of four components, with 87.61% of the accumulated variance explained, with several changes in the variables that contribute ($PM_{2.5}$, PM_{10}, PM_1, and O_3) to the components, about the results of the previous PCA. Object 6–1 (spring season) recorded lower values in its variables within the time-2 range, which caused the object to change by presenting a greater distance from the objects of the previous cluster (time-1 range). Two objects (8–2 and 13–2) in the summer, and similarly in the autumn (1–3 and 2–3), were identified that changed clusters due to a decrease in the concentration levels of indoor air pollutants recorded in that microenvironment and in that time range. In the particular case of office 12 (represented by objects 12–1, 12–2, and 12–3), which is an open work area that during the spring and summer seasons has a large number of visitors in time-2, which is reflected in the change of cluster of objects 12–1 and 12–2, and the case of object 12–3 it changes to a cluster with objects that register a lower concentration of air pollutants. On the other hand, elements 10–1, 12–1, 13–1, 12–2, and 9–3 registered a higher average in the concentration of air pollutants in the time-2 range; for this reason, they present a smaller distance with the elements of cluster 1, as shown in Fig. 11.

As mentioned above, the exterior walls of building A are made of glass, and several microenvironments receive sunlight through-out the workday, but with a high impact during time-2. Still, it was not found to influence air pollution concentration levels in the spring and summer periods when the highest ambient air temperatures occur during the year. For example, objects 7–1 and 11–1 (offices that receive direct sunlight for most of the day) show a slight increase in the concentration in the time-2 range in the spring season (see Fig. 11).

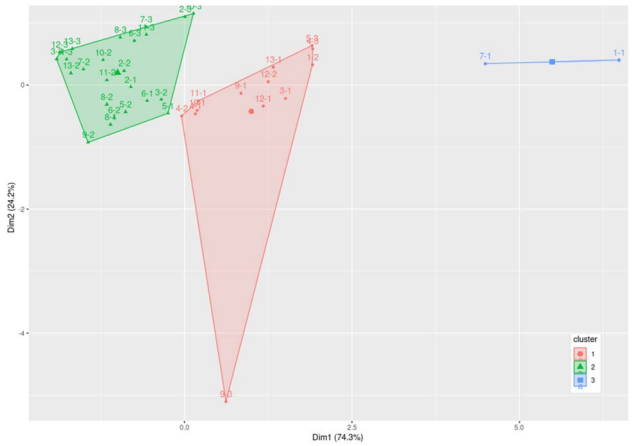

Fig. 9 Objects comprising the three clusters derived from the instances between 1:00 p.m. and 5:00 p.m., considering the result of the applied PCA method (building A).

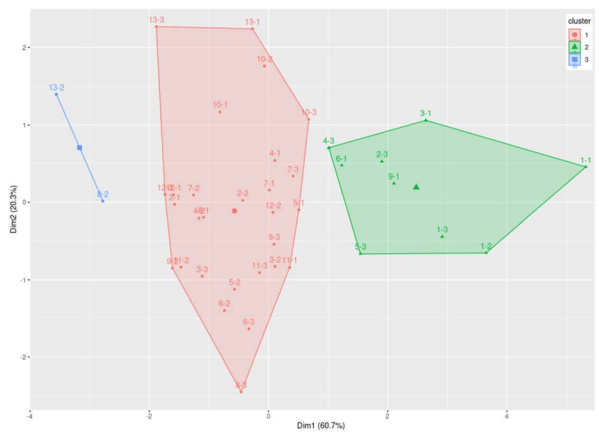

Fig. 10 Clusters formed with the *K*-means algorithm considering the instances of time-1 range and the variables of the result of the applied PCA (building A).

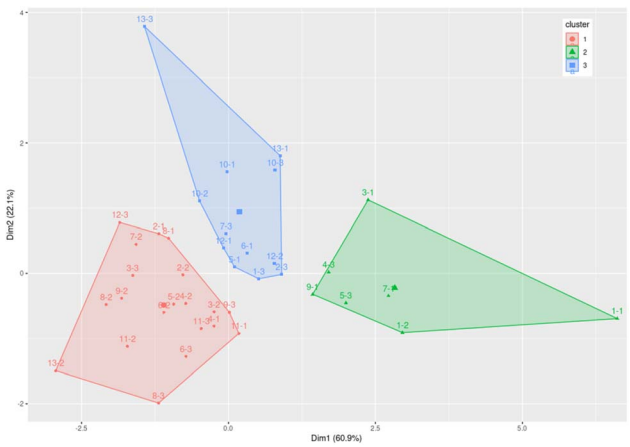

Fig. 11 Groups generated with data from building A, considering the variables derived from PCA analysis (four components).

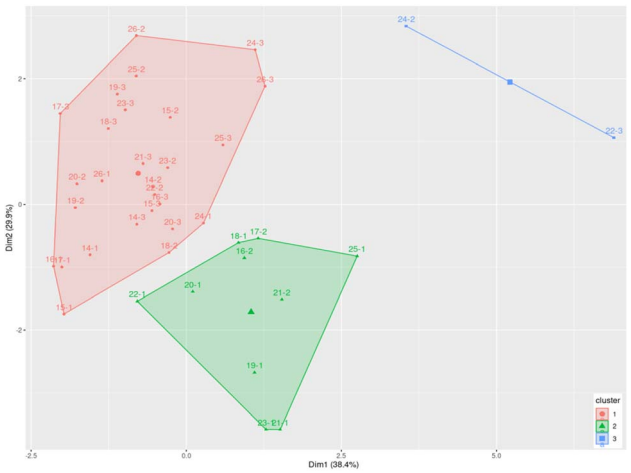

Fig. 12 Clusters generated from the data of building B, considering the 8 study variables.

In the case of the clusters generated from the data of building B, three groups of 10, 27, and 2 objects were formed (see Fig. 12). Cluster 2 comprises 6 elements from the spring period, 9 from the summer period, and 12 objects (92%) from the autumn period. The object from autumn that is not included corresponds to office 22–3, characterized by high levels of particulate matter in this period. Therefore, the distance of element 22–3 from the centroid and the objects of cluster 2 is considerable. On the contrary, the 12 elements identified in the autumn period (cluster 2) of building two that follow the same pattern show high similarities. This cluster comprises 6 and 9 objects from spring and summer, confirming that the concentration levels of indoor air pollutants and meteorological parameters are very similar in this building. On the other hand, cluster 3 comprises only two objects (22–3 and 24–2), which present high concentration levels in the three fractions of particulate matter.

In the second experiment, we are considering the sub-dataset of building B composed of $PM_{2.5}$, PM_{10}, PM_1, O_3 and CO_2 variables, assigned according to the results of the PCA, which is com-posed of 4 components with a proportion of explained variance of 90.33. The objects 24–1, 24–3, 25–3, and 26–3, which belonged to group 1 in the previous analysis (see Fig. 12), were reassigned to a different group (see Fig. 13). Furthermore, objects 20–1 and 22–1 were reassigned to cluster 2 when applying the PCA, while cluster 3 retained its original two elements. The rest of the elements (33) confirmed the same degree of similarity using a smaller number of variables in the sub-dataset, confirming the cohesion between the data reflected in the formation of the clusters.

In experiment 3 of building B, the PCA was applied to the time-1 and time-2 sub-datasets, using fewer components and variables to contribute to the dimensionality reduction. In time-1, three principal components are used with a proportion of explained variance of 80.80%, with a contribution from the following variables: $PM_{2.5}$, PM_{10}, PM_1, and CO_2. The above allows generating 3 clusters with 1, 32, and 6 objects, respectively. In time-2, the first four components were selected, which accumulated a variance of 81.87%. This experiment utilizes only three variables (PM_1, $PM_{2.5}$, and CO_2), which contribute the highest percentages to the selected principal components, enabling the construction of three clusters with 1, 12, and 26 objects. In this experiment, numerous changes are observed in the objects that comprise the clusters within the time-1 and time-2 ranges. Cluster 2, which is made up of 31 elements in the time-1 range (see Fig. 14), maintains only 26 objects (cluster 3) in the time-2 range (see Fig. 15). Now, their distance is smaller, approximately equal to the centroid of cluster 3, thereby maintaining a statistical similarity between the objects based on their data characteristics. In the case of cluster 2 (time-2), it incorporated 1 and 6 elements from cluster 1 and cluster 2 in time-1, respectively, which increased the average values in their variables in the time-2 range (see Fig. 15). That is, the distance between the new objects comprising cluster 2 is minimal, maintaining cohesion in the group. Finally, cluster 1 (time-2) incorporates an element that belonged to cluster 3 (time-1) and loses object 22–3, which considerably decreases the value of its characteristics in the time-2 range (see Fig. 14 and 15).

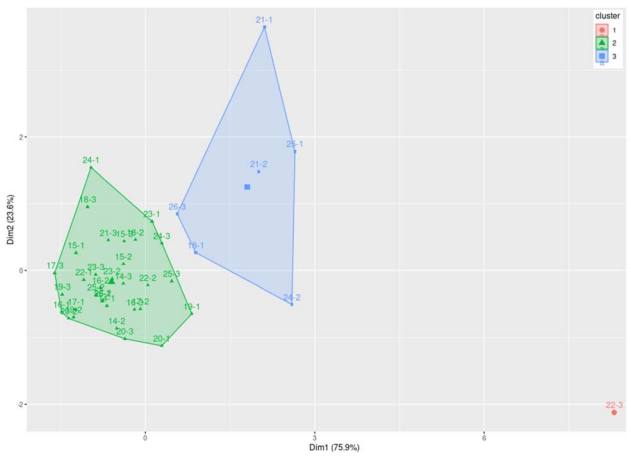

Fig. 13 Clusters formed considering instances of the time-1 range (building B).

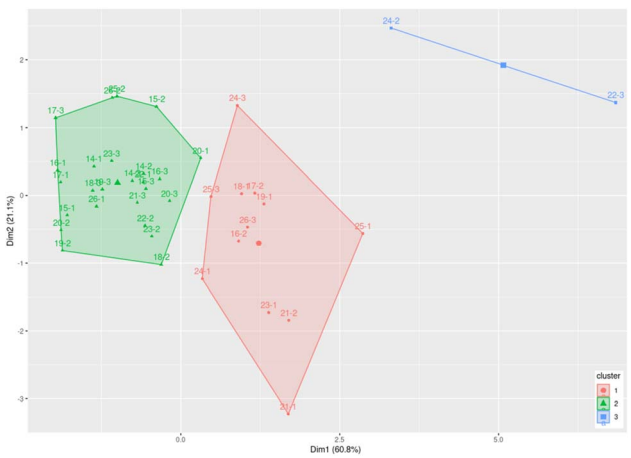

Fig. 14 Clusters generated from the PCA with data from building B.

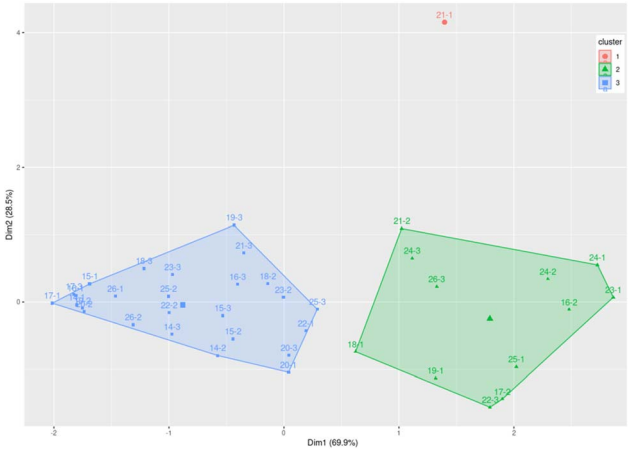

Fig. 15 Objects that comprise the clusters of building B derived from instances of time-2 range.

To compare the performance of the *K*-means clustering models, the DBSCAN and hierarchical (a Ward linkage method) algorithms were implemented using the datasets generated for the second experiment, where variables were selected through the PCA technique, ensuring highly correlated variables. The quality of the clusters produced by each algorithm and for each dataset was evaluated using the Davies–Bouldin Index (DBI) and the Silhouette score. It is essential to clarify that the two metrics offer slightly different perspectives; DBI emphasizes the relationship between internal dispersion and the distance between cluster centroids, while the Silhouette score emphasizes cohesion and separation at the level of each data point. As a result, the Silhouette score may be more sensitive to outliers or irregularly shaped clusters. Table 1 shows the results obtained in each experiment. In the clustering evaluation of the building A dataset, the hierarchical algorithm received the lowest score, indicating that its clusters are the densest and most distinctly separated compared to the other two algorithms. The performance of the *K*-means algorithm is very similar to that of hierarchical clustering, with its score suggesting that its clusters are less tight than those of the hierarchical model. Conversely, in the evaluation using the building B dataset, the DBScan algorithm scored the lowest, implying that its clusters are the most well-defined, offering the best balance of internal density and external separation (see Table 1). The *K*-means algorithm achieved the highest score, indicating that its clusters are the least compact or most overlapping compared to the other models.

Table 1 Values achieved in the quality assessment metrics of the constructed clusters

Dataset	Algorithm	BDI	Silhouette
Building A	*K*-means	1.3842	0.2559
	DBScan	1.3931	0.1472
	Hierarchical	1.3580	0.2515
Building B	*K*-means	1.0902	0.3792
	DBScan	0.8449	0.3812
	Hierarchical	0.9668	0.3510

On the other hand, the *K*-means algorithm achieved a relatively low positive Silhouette score of 0.2559, indicating that the clusters formed by *K*-means have weak separation and cohesion. Although no points are misclassified, the proximity between clusters is evident, showing that the boundaries are not well-defined. The values in Table 1 for the Silhouette score metric on the building A dataset, all of which are below 0.3, suggest that the dataset may lack a natural and clearly defined cluster structure. Regarding the quality of the clusters built with the building B dataset, the DBScan and *K*-means algorithms achieve very similar metrics (the difference from DBScan is only 0.002), which can be considered comparable in practice. This result indicates that they have produced the most well-defined clusters with the best separation compared to the hierarchical algorithm. Fig. 16 shows the silhouette coefficient of the clusters obtained with the dataset from the time-2 range of building B. This metric measures the quality of the generated clusters. The coefficient obtained is 0.52, indicating that the groups are well separated and differentiated, with good cohesion within each cluster and distinct separation between clusters.

Additionally, the stability of the clusters was statistically assessed over different time intervals using the Adjusted Rand Index (ARI) to ensure that the clusters are not the result of random fluctuations in the data. These intervals correspond to the spring, summer, and fall data partitions. The ARI measures the similarity between two clusters, adjusting for the likelihood of agreement by chance. An ARI of 1 indicates perfect agreement, while a value of 0 or less suggests that the similarity is what would be expected by chance. The *K*-means algorithm was applied to each partition with the same number of clusters. Table 2 shows the ARI values for the pairs of partitions within each building. The ARI values for the three partition pairs in building A are very similar. The ARI of 0.5687204 between the spring and summer clusters indicates moderate agreement, confirming a strong correlation between the cluster structures in spring and summer. These values suggest that the cluster structure remains somewhat stable over time but is not entirely static. Some data points may have shifted clusters, or the boundaries between clusters may have shifted due to seasonal changes. On the other hand, assessing the similarity between the spring and autumn partitions of building B yields an index value of 0.6381156, indicating that the cluster structure is highly similar between spring and autumn (see Table 2). This points to significant stability in the data patterns over time, despite possible seasonal changes. Similarly, the similarity index between the spring and summer partitions was 0.6042155, confirming that the clustering algorithm produced a meaningful partition that closely matches the reference. While this is not a perfect agreement, it is a positive outcome showing that the algorithm captured a real and relevant data structure.

Furthermore, the loading values of each variable in PCA were analyzed to quantify intraday changes using a sliding window (time-1 and time-2) to determine the significance of the variables. A loading value is the correlation between an original variable and a principal component, where a high loading

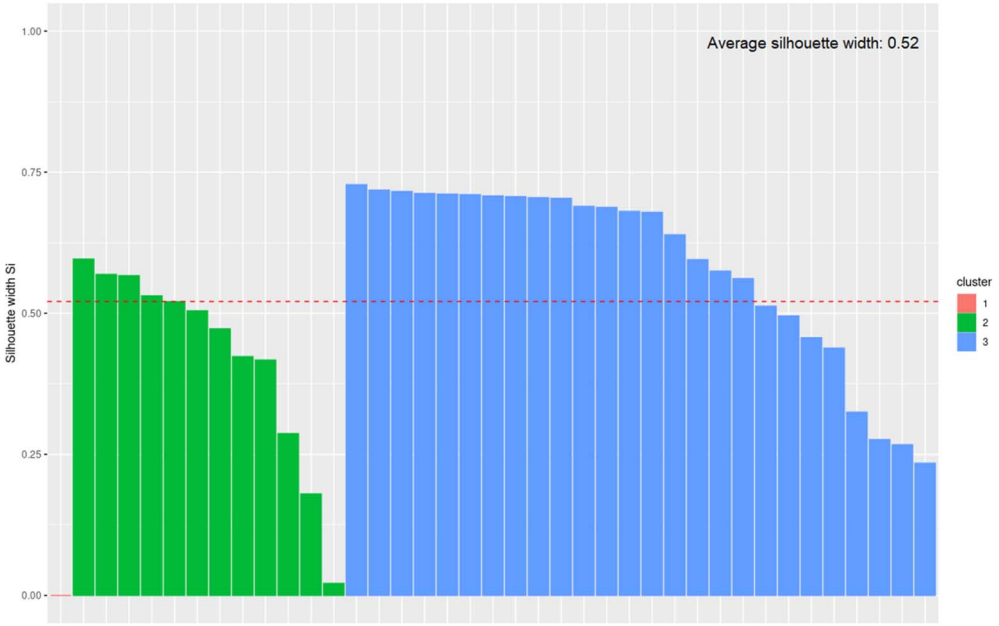

Fig. 16 Silhouette score of the clusters generated for the time-2 range of building B.

(positive or negative) on the first principal component (PC1) indicates that the variable is a key factor influencing data variance in that time window. In building A, the loading values of the PM_1 variable on PC1 reflect its contribution to data variability in that component. With a loading value of 0.53508485 in the morning and 0.531226037 in the afternoon, it suggests that the variable's importance is comparable and high during both periods, with a slight decrease in the afternoon (see Table 3). Hence, the PM_1, $PM_{2.5}$, and PM_{10} variables are major factors in the data variability in both morning and afternoon; this indicates that the data structure captured by PCA remains relatively stable concerning the influence of these variables. The O_3 variable, with a loading value of about 0.095, has a very low impact on the main pattern of data variability during the morning. This is expected, as tropospheric ozone production is linked to solar radiation. In the second time window (afternoon hours), the loading value rises to 0.163, which is a 71% increase compared to the morning (see Table 3). Although this value remains moderate, the increase is significant and indicates that ozone concentration variability be-comes more important in the afternoon data, aligning with its photochemical formation cycle. For its part, the variable CO is not a significant factor in explaining the data's variability in either period. This indicates that CO fluctuations are independent of the other factors influencing PC1.

Concerning CO_2, the difference between the loadings shows that its relative importance varies considerably throughout the day. Its influence is moderate in the morning but nearly zero in the afternoon. Regarding the meteorological variables, the difference in loadings for relative humidity shows that its importance shifts significantly throughout the day. While it plays a moderate role in the morning (0.163), it becomes a key factor in the afternoon (0.313). This change may be linked to the use of air conditioning, the influx of people, or the rising temperature during the day, making relative humidity a more representative indicator of the overall environmental variability during that period (see Table 3). In this sense, the difference in temperature loads shows that its behavior and influence change significantly throughout the day. In the morning, temperature exhibits a behavior that positively contributes to the main pattern of variability (0.139). In the afternoon, their relationship reverses, and their relative influence decreases (−0.080). This pattern could be related to external factors like solar radiation

Table 2 Results of the cluster stability assessment using ARI

Partition combination	Building A	Building B
Spring – summer	0.5687204	0.6042155
Spring – autumn	0.5316253	0.6381156
Summer – autumn	0.5433255	0.4400574

Table 3 First principal component (PC1) loadings by time period

Dataset	Variable	Time-1 loadings	Time-2 loadings
Building A	PM_1	0.53508485	0.53122603
	$PM_{2.5}$	0.54423959	0.54011290
	PM_{10}	0.53900124	0.53932743
	O_3	0.09493903	0.16263629
	CO	0.02258891	−0.00078407
	CO_2	−0.26714783	−0.06874059
	RH	0.16330848	0.31253203
	T	0.13929765	−0.07968783
Building B	PM_1	0.54792026	0.4602331
	$PM_{2.5}$	0.55347504	0.4069517
	PM_{10}	0.5536367	0.1085071
	O_3	0.13994936	−0.3379548
	CO	0.08678465	−0.2711114
	CO_2	0.11967097	0.2571383
	RH	0.18287404	−0.3403311
	T	−0.10978945	0.4910619

or the intervention of air conditioning systems, which alter the natural behavior of temperature in the office environment.

Otherwise, the analysis of PCA loadings in building B revealed that PM_1 and $PM_{2.5}$ are key contributors to data variability in-side the building during both periods, with their influence being stronger in the morning. For PM_1, the differences in loading values indicate that its relative importance among the variables changes throughout the day (see Table 3). This decrease can result from several factors in an office environment, such as reduced outdoor PM_1 and $PM_{2.5}$ input due to less traffic or external activities in the afternoon; more intense indoor activities that generate PM_1 and $PM_{2.5}$ in the morning and decline later; and other factors like temperature, relative humidity, or other pollutants becoming more dominant in the afternoon, making PM's variability less influential on PC1. Additionally, the high correlation between these two variables suggests that the morning PCA may be capturing the dynamics of outdoor air pollution entering the building. Likewise, the change in the PCA loadings of the PM_{10} variable, from 0.553 to 0.108, indicates that its contribution and significance to the building's indoor air variability patterns decreased significantly from morning to afternoon (see Table 3). In the afternoon, the building's windows may have been closed, or external conditions such as traffic may have changed. This greatly reduces the entry of outdoor PM_{10}, removing the main source of variability in the morning. Additionally, there is a shift in pollution sources, becoming mainly indoor during the second period, when reduced building operations can decrease dust resuspended by human activity. Therefore, the PM_{10} concentration in the afternoon either behaves independently or is affected by local and sporadic factors that do not contribute to the overall variance explained by the PCA.

The ozone behavior in building B mirrors that in building A. The afternoon increase in O_3 load suggests that ozone shifts from being a minor variable to one that accounts for a large part of the indoor environment's variability. This is likely caused by increased outdoor ozone infiltration during the afternoon, due to more sunlight, higher outdoor O_3 levels, and the building's natural or mechanical ventilation. The change in PCA loadings for the CO variable shows it rising from an insignificant role in the morning to a significant one in the afternoon. The negative sign indicates an inverse relationship, probably related to CO dilution through natural ventilation or external source dynamics. Likewise, the increase in PCA loadings for CO_2 from 0.119 in the morning to 0.257 in the afternoon is a typical and reasonable trend. It shows that CO_2 becomes much more influential in explaining indoor air variability as the day goes on.

Relative humidity became much more significant for the principal component in the afternoon period (the absolute value of the loading increased to −0.34), indicating that its fluctuations are now a key factor in explaining indoor environment variability. The negative loading shows an inverse relationship, meaning that as the principal component increases, relative humidity decreases. In conclusion, during the afternoon, as indoor temperature rises (due to solar heating or human activity), relative humidity drops (because warm air can hold more water vapor before becoming saturated). Therefore, the afternoon principal component might be capturing the building's heating pattern and the resulting decrease in relative humidity. In this way, the temperature variable shows a change in PCA loadings from −0.109 in the morning to 0.491 in the afternoon, indicating that temperature becomes a significant and dominant factor in indoor air dynamics as the day progresses. Furthermore, during the afternoon, the sun shines directly on the building and its windows, causing a natural rise in interior temperature. In this context, as the temperature rises, it can affect other factors such as relative humidity (which decreases) and human activity (which can increase or de-crease thermal comfort). Furthermore, using electronic devices or the presence of more people in the afternoon can generate heat, contributing to the increase in temperature.

Finally, a Spearman correlation analysis was performed between the study variables and contextual and operational variables in each microenvironment to identify activities that may influence the levels of air pollution recorded in the two buildings. In this analysis, the operation of printers and copiers, computers, electronic devices (such as IP phones, access points, and external storage units), and mini-refrigerators was recorded, along with additional activities like vacuum cleaning and electronic soldering. A daily average of door openings and closings was calculated for each microenvironment. Table 4 shows an excerpt of the correlations with a significance value of $p > 0.05$ calculated for the spring period in building A. A strong correlation is observed between particulate matter fractions and printers, with a coefficient greater than 0.77. Additionally, computer operation and door opening/closing have coefficients ranging from 0.55 to 0.64. This suggests that when equipment is in use and people enter or leave the office, particulate matter concentrations tend to increase.

Table 5 displays the correlation coefficient matrix for autumn data in building B, revealing a similar pattern between particulate matter and printers, computers, and door movements, with the correlation for door activity being notably higher (0.87, p-value < 0.001). Moderate to strong relationships were also found between O_3 and printers (0.50), devices (0.62), and operational activities (0.62), indicating that increased equipment use and activities such as vacuuming or soldering electronic components lead to higher O_3 levels. Additionally, RH shows a negative association with printers and a positive link to door opening/closing, implying that printer operation tends to decrease relative humidity in the microenvironment,

Table 4 Spearman correlation coefficient matrix for the spring season in building A[a]

	Printers	PCs	Devices	Doors	Activities
PM_1	0.86***	0.57*	—	0.55*	—
$PM_{2.5}$	0.78**	0.57*	—	0.64*	—
PM_{10}	0.78**	0.57*	—	0.62*	—
T	—	—	—	—	−0.58*

[a] Where * equals $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

Table 5 Spearman correlation coefficient matrix for the autumn season in building B[a]

	Printers	PCs	Devices	Doors	Activities
PM_1	0.89***	0.60*	—	0.87***	—
$PM_{2.5}$	0.89***	0.60*	—	0.87***	—
PM_{10}	0.89***	0.60*	—	0.87***	—
O_3	0.50*	—	0.62*	—	0.62*
RH	−0.54*	—	—	0.77**	—

[a] Where * equals $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

while opening a door causes humidity levels to rise due to cold air entering from the corridors.

4. Discussion

The following conclusions can be drawn from evaluating the quality of the *K*-means clustering models. Clusters formed by *K*-means using PCA-filtered and divided by time range datasets exhibit non-convex shapes. In this sense, DBI may be less sensitive to clusters with non-convex shapes or highly variable densities, as its calculation is based on distances to centroids. For this reason, the values obtained for *K*-means clustering quality metrics are not the most appropriate; algorithms such as DBScan achieve better results due to their ability to handle this kind of data and clusters with elongated or curved shapes (see Table 1). This is shown in Fig. 9 and 13, as well as in the Silhouette score values obtained by the clustering algorithms, where *K*-means produces the best scores for clusters with convex shapes and a low quality for clusters with elongated or linear shapes.

The cluster analysis enabled the identification of microenvironments that maintain similar values in the variables across the three seasons considered in the study. The 20 objects (microenvironments) that compose cluster 1 of building A (see Fig. 9) are characterized by low values in the average concentration of particulate matter, in $PM_{2.5}$ from 1.5 to 9.2 μg m^{-3} and in PM_{10} from 1.7 to 10.3 μg m^{-3}. In all the objects of this cluster, high levels of CO concentration are present, exceeding 0.90 ppm, which stands out as the highest records, considering the data from the three seasons of building A. In addition, 50% of the objects that comprise cluster 1 recorded the highest CO_2 concentration values. The patterns described above define the maximum similarity between these objects. On the other hand, the objects of cluster 2 stand out for having low values of CO, between 0.63 ppm and 0.86 ppm, and low concentrations of PM_1, $PM_{2.5}$, and PM_{10}. The variables of the particulate matter fractions present concentration levels very similar to those of cluster 1. However, the separation between these clusters is defined by the difference in CO and CO_2 concentration levels, which are lower than those recorded in cluster 1. In the case of cluster 3, the values of the concentration levels of PM_1, $PM_{2.5}$, and PM_{10} are higher than those recorded in the objects of clusters 1 and 2; for example, PM_{10} has concentrations between 15 and 22 μg m^{-3}, which causes the separation of these objects to form a third cluster.

In this sense, the objects 10–1, 10–2, and 10–3 in cluster 3 (see Fig. 9), which correspond to office 10 in the spring, summer, and autumn seasons, are most similar. Likewise, objects 12–1 and 12–2, as well as elements 13–1 and 13–3, are highly similar to the other objects that comprise the cluster. The coincidence of objects representing the same office but from different seasons of the year allows us to conclude that the activities carried out in these microenvironments do not vary in frequency or intensity. In addition, the sources of air pollutants released at the same concentration level are probably the same sources that originate the pollution in the 3 time periods.

In the clusters derived from the data collected from building B (see Fig. 13), it is evident that the objects that form cluster 3 are less similar or dissimilar to the objects of other clusters, and there is a considerable distance between the objects of cluster 3 and the objects of clusters 1 and 2. This behavior in the objects of cluster 3 is caused by the very high concentration levels of PM_1, $PM_{2.5}$, and PM_{10}, from 25 to 42 μg m^{-3}, 36 to 66 μg m^{-3}, and 44 to 79 μg m^{-3}, respectively; these levels of contamination are very far from the data recorded in the other objects. It is essential to note that the IT technical support activities frequently carried out in these offices often result in high contamination levels. Computer equipment maintenance activities take place in this office, which can cause dust to spread from the vacuum cleaners used for cleaning the equipment. Electronic components are also occasionally soldered, releasing polluting gases into the interior, even when an extractor fan is installed. Seven people work in this area, following the same schedule. During the spring, the door was opened 42 times in a single workday. This door grants direct access to the exterior of the building. Therefore, the IT support activities performed in these offices may be related to the high levels of pollution recorded.

Furthermore, the composition of clusters 2 and 3 is determined by the concentration levels of the particulate matter fractions. In the case of cluster 1, which consists of 12 objects, PM_1 concentration levels were recorded between 14 and 20 μg m^{-3}, $PM_{2.5}$ with values between 16 and 36 μg m^{-3}, and PM_{10} values between 17.5 and 41 μg m^{-3} were observed. In cluster 2, $PM_{2.5}$ concentration levels are between 2.80 and 16 μg m^{-3} and PM_{10} between 3.3 and 21.6 μg m^{-3}. In the case of CO, variations are observed in the three clusters. Similarly, the concentration of CO_2 is only greater than 928 ppm in cluster 1, where six objects are observed.

On the other hand, the average values of CO in the indoor air of the buildings considered in the study did not exceed the values defined in the guidelines of the US Environmental Protection Agency (EPA) of 9 ppm,[54] and by the World Health Organization (WHO) of 8.6 ppm for pollutants in the IAQ.[55] The mean concentration of the CO of each building is consistent with that reported by Baloch *et al.*,[4] where the distribution of indoor air pollutants in approximately 300 classrooms across various countries. The CO concentration was evaluated through a 30 minutes short-term measurement in each room, yielding a mean CO concentration of 0.72 ppm. On the other hand, in Moreira *et al.*,[56] an average of 1.8 ppm is reported for the pollutant CO, with conditions like our experiment in

temperature and relative humidity; in a study carried out in 3 offices and four service areas in a school involved teachers, workers, and students. Similarly, Shen *et al.*,[5] reports the average levels of CO between 1.2 and 1.6 ppm inside different rooms in regular residential apartments.

The mean concentrations of CO_2 for the two buildings ranged from 634 to 772 ppm. It is essential to note that the average CO_2 concentration in the three seasons of the two buildings exceeds 600 ppm.[12] However, although the averages do not exceed the limits established by the EPA, maximum values above 2000 can be observed in building A during the summer (2423 ppm) and in building B during spring and autumn, at 2715 and 2015 ppm, respectively. In particular, in building B, the average CO_2 concentration exceeded 770 ppm during the spring monitoring season. Therefore, a person's health impact cannot be ruled out in this building, as the average CO_2 level is very close to the permissible limit value in IAQ during the seasons analyzed in this study. Building B has the peaks (2715 ppm, 2015 ppm, and 1463 ppm) and average highest CO_2 concentrations. In the spring, the highest mean measurements of our experiment are observed (772 ppm $\pm$ 428 ppm), which can be influenced by the very high ambient temperature during this season. With a sudden and extreme change in temperature between winter and spring, a maximum temperature of 28 °C in the winter last weeks and spring with a maximum temperature of 40 °C (in the range from 1:00 p.m. to 7:00 p.m.). Hence, the air conditioning system of the building is in operation for a minimum of 12 hours per day, and all the building windows are closed permanently. On the other hand, during the autumn period, the average CO_2 concentration decreased by approximately 50 ppm (from 716 ppm), which is consistent with the decrease in ambient temperature. It is common to open the windows in buildings during these seasons of the year.

It is essential to consider that several offices of building B are staff areas, and there is an influx of students at certain times of the day. Then, the average CO_2 measurements and the high peaks observed can be caused by the increase in people in the indoor spaces of this building. In this sense, Szabados *et al.*,[57] exposed an average CO_2 of 1329 ppm, with a minimum concentration value of 767 ppm and a maximum of 2328 ppm. These data are derived from a study in which 64 school buildings were monitored for five consecutive days, with periods of 6 to 8 hours per day. Similarly, Gupta *et al.*,[58] reported monitoring IAQ in four office buildings during eight regular hours of office work on weekdays; the average CO_2 concentration in some buildings was 1434 ppm, but with very high average concentrations in others (1918 ppm). In this way, in Villanueva *et al.*,[59] high peak and average CO_2 concentrations have been reported in secondary school classrooms during the reopening after the COVID-19 pandemic. The average concentration of CO_2 was 699 ppm ($\pm$172 ppm), with a minimum value of 393 ppm and a maximum of 2117 ppm, based on data collected through continuous monitoring for approximately one month, 6 hours per day.

In addition, in Madureira *et al.*,[60] present a study of the exposure of newborns and mothers indoors in northern Portugal, identifying high concentrations of $PM_{2.5}$ and PM_{10} with average levels of 53 $\mu g\ m^{-3}$ and 57 $\mu g\ m^{-3}$ in a sample of 65 homes. Furthermore, Qiu *et al.*,[61] found a significant variation in $PM_{2.5}$ concentration levels between floors of tall buildings with average values (24 hours) between 34 and 102 $\mu g\ m^{-3}$. On the other hand, Roh *et al.*,[62] reported average values in $PM_{2.5}$ concentration between 4.28 $\mu g\ m^{-3}$ and 12.2 $\mu g\ m^{-3}$ in a study of 8 offices before the COVID-19 pandemic. Similarly, in a survey of indoor air quality in 25 offices of a Medical University,[63] an average of 21 $\mu g\ m^{-3}$ was recorded with a minimum of 3 $\mu g\ m^{-3}$ and a maximum of 65 $\mu g\ m^{-3}$, considering monitoring of 2 continuous hours per office. Finally, in Felgueiras *et al.*,[64] report in a study that considers 15 offices, they found high concentration levels of $PM_{2.5}$ and PM_{10} only in one office located on a second floor, with values of 53 $\mu g\ m^{-3}$ and 57 $\mu g\ m^{-3}$, respectively; in the rest of the offices, values less than 20 $\mu g\ m^{-3}$ and 40 $\mu g\ m^{-3}$ were recorded for $PM_{2.5}$ and PM_{10}.

In summary, the CO, CO_2, O_3, $PM_{2.5}$, and PM_{10} levels recorded in this study are within a normal concentration range for each pollutant. The PM_{10} average concentration for 24 hours does not exceed the permissible limit of 45 $\mu g\ m^{-3}$ published by the WHO.[65] In the case of $PM_{2.5}$, the allowable limit of 15 $\mu g\ m^{-3}$ is slightly exceeded in building B in the spring and summer with a value less than 1 $\mu g\ m^{-3}$, and in the fall with a value less than two $\mu g\ m^{-3}$, which corresponds to 12.9%. The average CO, CO_2, and O_3 concentrations recorded in the two buildings are far from the maximum limits recommended by different international organizations. For example, CO_2 is found at 23% and 36% of the recommended limit of 1000 ppm, respectively.

5 Conclusions

The proposed clustering approach, based on the *K*-means algorithm, can effectively improve the identification of microenvironment clusters by leveraging patterns discovered in their characteristics. This study enables the creation of maps of microenvironments based on the objects that comprise the clusters. The microenvironments (objects) within each cluster are more comparable to one another and distinct from the microenvironments of the other groups generated using the *K*-means algorithm. Furthermore, the implementation of the principal component analysis (PCA) method allowed us to identify the variables with the most significant contribution to the creation of groups, thus detecting the variables that influence the creation of the groups, highlighting that in some cases, the clusters represent offices with a low, medium, or high level of pollution. Moreover, in the analysis by time range, it was observed that the O_3 pollutant variable makes a significant contribution to the formation of the clusters, particularly when solar radiation is high and ambient temperatures are also high.

Author contributions

Conceptualization: BAM-H, ET-L, JMJ-P, RV-H. Data curation: ET-L. Formal analysis: BAM-H, ET-L. Funding acquisition: BAM-H, ET-L. Investigation: BAM-H, ET-L, JMJ-P, RV-H. Methodology: ET-L. Project administration: BAM-H, ET-L. Resources:

BAM-H, ET-L. Software: ET-L. Supervision: BAM-H. Validation: ET-L. Visualization: JMJ-P, ET-L. Writing – original draft: BAM-H, ET-L, JMJ-P, RV-H. Writing – review & editing: BAM-H, ET-L.

Conflicts of interest

There are no conflicts to declare.

Data availability

Data for this article, including air pollution concentrations (PM_1, $PM_{2.5}$, PM_{10}, CO, CO_2, and O_3) and meteorological parameters (temperature and relative humidity), are available at "Dataset of indoor air pollutants at the microenvironment level in buildings within a university center" in the Mendeley Data repository at **https://doi.org/10.17632/tx3v4kyfsw.1**.

Acknowledgements

The Autonomous University of Tamaulipas partially funded this research. Additionally, the study received partial funding from the Secretariat of Science, Humanities, Technology, and Innovation (SECIHTI) through grant 1239803 (Jailene Marlen Jaramillo-Perez).

References

1 L. Zaniboni and R. Albatici, *Buildings*, 2022, **12**, 1–10.
2 J. Wu, J. Weng, B. Xia, Y. Zhao and Q. Song, *Int. J. Environ. Res. Public Health*, 2021, **18**, 1–10.
3 I. L. Niza, A. M. Bueno, M. Gameiro da Silva and E. E. Broday, *Results Eng.*, 2024, **24**, 103157.
4 R. M. Baloch, C. N. Maesano, J. Christoffersen, S. Banerjee, M. Gabriel, *et al.*, *Sci. Total Environ.*, 2020, **739**, 139870.
5 G. Shen, S. Ainiwaer, Y. Zhu, S. Zheng, W. Hou, H. Shen, Y. Chen, X. Wang, H. Cheng and S. Tao, *Environ. Pollut.*, 2020, **267**, 115493.
6 A. Chamseddine, I. Alameddine, M. Hatzopoulou and M. El-Fadel, *Build. Environ.*, 2019, **148**, 689–700.
7 L. Schibuola and C. Tambani, *Atmos. Pollut. Res.*, 2020, **11**, 332–342.
8 V. Sahu and B. R. Gurjar, *Build. Environ.*, 2020, **185**, 107310.
9 M. Mannan and S. G. Al-Ghamdi, *Int. J. Environ. Res. Public Health*, 2021, **18**, 1–25.
10 Environmental Protection Agency, *Why indoor air quality is important to schools*, **https://www.epa.gov/iaq-schools/whyindoor-air-quality-important-schools**, 2020.
11 H. Zhang, D. Yang, V. W. Tam, Y. Tao, G. Zhang, S. Setunge and L. Shi, *Renewable Sustainable Energy Rev.*, 2021, **141**, 110795.
12 Occupational Safety Health Administration, *OSHA Technical Manual Section 453 III: Chapter 2: Indoor Air Quality Investigation*, U.S. department of labor technical report, 2017.
13 A.-R. Ali, A. Rasheed, H. Abdelmaoula and A.-R. Yasmin, *Adv. Meteorol.*, 2021, 6680476.
14 H. Saidin, A. A. Razak, M. F. Mohamad, A. Z. Ul-Saufie, S. A. Zaki and N. Othman, *Buildings*, 2023, **13**, 1–18.
15 A. Asif and M. Zeeshan, *J. Build. Eng.*, 2023, **72**, 106687.
16 W.-Y. Aung, M. Noguchi, E.-E. Pan-Nu Yi, Z. Thant, S. Uchiyama, T.-T. Win-Shwe, N. Kunugita and O. Mar, *Atmos. Pollut. Res.*, 2019, **10**, 722–730.
17 F. J. Kelly and J. C. Fussell, *Atmos. Environ.*, 2019, **200**, 90–109.
18 L. Qabbal, Z. Younsi and H. Naji, *Indoor Built Environ.*, 2022, **31**, 586–606.
19 N. Ma, D. Aviv, H. Guo and W. W. Braham, *Renewable Sustainable Energy Rev.*, 2021, **135**, 110436.
20 P. Kapalo, S. Vilcekova, L. Meciarova, F. Domnita and M. Adamski, *Sustainability*, 2020, **12**, 1–10.
21 Environmental Protection Agency, *Introduction to Indoor Air Quality*, **https://www.epa.gov/indoor-air-qualityiaq/introduction-indoor-air-quality**, 2024, Accessed: 26.08.2024.
22 S. S. Korsavi, A. Montazami and D. Mumovic, *Indoor Air*, 2021, **31**, 480–501.
23 S. Derek G., G. Lauren N., P. Joseph A. and M. Jason, *J. Environ. Public Health*, 2021, 5580616.
24 E. R. Jones, J. G. Cedeño Laurent, A. S. Young, P. MacNaughton, B. A. Coull, J. D. Spengler and J. G. Allen, *Build. Environ.*, 2021, **200**, 107975.
25 H. Wang, X. Guo, J. Yang, Z. Gao, M. Zhang and F. Xu, *J. Build. Eng.*, 2024, **95**, 110168.
26 W. Hou, J. Wang, R. Hu, Y. Chen, J. Shi, X. Lin, Y. Qin, P. Zhang, W. Du and S. Tao, *Environ. Int.*, 2024, **186**, 108641.
27 M. S. Hussain, G. Gupta, R. Mishra, N. Patel, S. Gupta, S. I. Alzarea, I. Kazmi, P. Kumbhar, J. Disouza, H. Dureja, N. Kukreti, S. K. Singh and K. Dua, *Pathol., Res. Pract.*, 2024, **255**, 155157.
28 C. Shrubsole, S. Dimitroulopoulou, K. Foxall, B. Gadeberg and A. Doutsi, *Build. Environ.*, 2019, **165**, 106382.
29 Z. Liu, G. Wang, L. Zhao and G. Yang, *IEEE Access*, 2021, **9**, 70479–70492.
30 V. V. Tran, D. Park and Y.-C. Lee, *Int. J. Environ. Res. Public Health*, 2020, **17**, 1–27.
31 M. Diaz, M. Cools, M. Trebilcock, B. Piderit-Moreno and S. Attia, *Sustainability*, 2021, **13**, 1–16.
32 K. B. Shah, D. Kim, S. D. Pinakana, M. Hobosyan, A. Montes and A. U. Raysoni, *Environments*, 2024, **11**, 1–14.
33 Y. Huang, Z. Yang and Z. Gao, *Int. J. Environ. Res. Public Health*, 2019, **16**, 1–16.
34 W. W. Nazaroff and C. J. Weschler, *Indoor Air*, 2022, **32**, e12942.
35 S. Zhao, Y. Yu, D. Qin, D. Yin, L. Dong and J. He, *Atmos. Pollut. Res.*, 2019, **10**, 374–385.
36 R. Borge, D. Jung, I. Lejarraga, D. de la Paz and J. M. Cordero, *Atmos. Environ.*, 2022, **287**, 119258.
37 X. Sha, Z. Ma, S. Sethuvenkatraman and W. Li, *J. Build. Eng.*, 2023, **76**, 107289.
38 S. Miao, M. Gangolells and B. Tejedor, *Indoor Air*, 2025, **2025**, 4453536.
39 J. Cebolla-Alemany, M. Macarulla Martí, M. Viana, V. Moreno-Martín, V. San Félix and D. Bou, *Build. Environ.*, 2024, **266**, 112091.

40 A. Caron, N. Redon, P. Coddeville and B. Hanoune, *Sens. Actuators, B*, 2019, **297**, 126709.
41 C.-H. Hsu and F.-Y. Cheng, *Aerosol Air Qual. Res.*, 2019, **19**, 1139–1151.
42 L. Liu and F. Zheng, *Alexandria Eng. J.*, 2024, **105**, 204–217.
43 P. Govender and V. Sivakumar, *Atmos. Pollut. Res.*, 2020, **11**, 40–56.
44 S. Wang, Q. Li, C. Zhao, X. Zhu, H. Yuan and T. Dai, *Inf. Sci.*, 2021, **542**, 24–39.
45 A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija and J. Heming, *Inf. Sci.*, 2023, **622**, 178–210.
46 M. Ahmed, R. Seraj and S. M. S. Islam, *Electronics*, 2020, **9**, 1–12.
47 H. Hu, J. Liu, X. Zhang and M. Fang, *Pattern Recognit.*, 2023, **139**, 109404.
48 Libelium, *Smart Environment PRO - Waspmote Gases PRO v30 Board*, **https://development.libelium.com/air-qualitystation/docs/datasheet**, 2023, Accessed: 03.17.2024.
49 PLANTOWER, *PMS7003 particulate matter sensor*, **https://www.plantower.com/en/products_33/76.html**, 2024, Accessed: 21.05.2024.
50 BOSCH BME280, *Humidity sensor measuring relative humidity, barometric pressure and ambient temperature*, **https://www.bosch-sensortec.com/products/environmentalsensors/humidity-sensors-bme280/#technical**, 2024, Accessed: 21.05.2024.
51 R. Duvall, A. Clements, G. Hagler, A. Kamal, V. Kilaru, L. Goodman, S. Frederick, K. Johnson Barkjohn, I. VonWald, D. Greene and T. Dye, *Performance Testing Protocols, Metrics, and Target Values for Ozone Air Sensors: Use in Ambient, Outdoor, Fixed Site, Non-Regulatory and Informational Monitoring Applications- EPA/600/R-20/279*, 2021, **https://cfpub.epa.gov/si/si_public_record_Report.cfm?dirEntryId=350784&Lab=CEMM**.
52 R. Duvall, A. Clements, G. Hagler, A. Kamal, V. Kilaru, L. Goodman, S. Frederick, K. Johnson Barkjohn, I. VonWald, D. Greene and T. Dye, *Performance Testing Protocols, Metrics, and Target Values for Fine Particulate Matter Air Sensors: Use in Ambient, Outdoor, Fixed Site, Non-Regulatory Supplemental and Informational Monitoring Applications - EPA/600/R-20/280*, 2021, **https://cfpub.epa.gov/si/si_public_record_Report.cfm?LAB=CEMM&dirEntryID=350785**.
53 B. A. Macías-Hernández, E. Tello-Leal, O. Barrios S., M. A. Leiva-Guzmán and R. Toro A, *Urban Clim.*, 2023, **52**, 101753.
54 Environmental Protection Agency, *Typical indoor air pollutants*, **https://www.epa.gov/sites/production/files/2014-08/documents/refguide_appendix_e.pdf**, 2009.
55 World Health Organization, *WHO guidelines for indoor air quality: selected pollutants*, **https://www.who.int/publications/i/item/9789289002134**, 2010.
56 F. Moreira, A. Ferreira, J. P. Figueiredo and I. Caseiro, *Proceedings of the 1st International Conference on Water Energy Food and Sustainability*, ICoWEFS, Cham, 2021, pp. 526–536.
57 M. Szabados, Z. Csákó, B. Kotlík, H. Kazmarová, A. Kozajda, A. Jutraz, A. Kukec, P. Otorepec, A. Dongiovanni, A. Di Maggio, S. Fraire and T. Szigeti, *Indoor Air*, 2021, **31**, 989–1003.
58 A. Gupta, R. Goyal, P. Kulshreshtha and A. Jain, *Indoor Environ. Qual.*, 2020, 67–76.
59 F. Villanueva, A. Notario, B. Cabañas, P. Martín, S. Salgado and M. F. Gabriel, *Environ. Res.*, 2021, **197**, 111092.
60 J. Madureira, K. Slezakova, C. Costa, M. C. Pereira and J. P. Teixeira, *Environ. Pollut.*, 2020, **264**, 114746.
61 Y. Qiu, Y. Wang and Y. Tang, *Build. Simul.*, 2020, **13**, 1009–1020.
62 T. Roh, A. Moreno-Rangel, J. Baek, A. Obeng, N. T. Hasan and G. Carrillo, *Atmosphere*, 2021, **12**, 1–11.
63 V. Surawattanasakul, W. Sirikul, R. Sapbamrer, K. Wangsan, J. Panumasvivat, P. Assavanopakun and S. Muangkaew, *Int. J. Environ. Res. Public Health*, 2022, **19**, 1–14.
64 F. Felgueiras, Z. Mourão, A. Moreira and M. F. Gabriel, *Build. Environ.*, 2024, **254**, 111393.
65 W. H. Organization, *WHO Global Air Quality Guidelines: Particulate Matter (PM2.5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide*, 2021.