INDOOR AIR POLLUTION ESTIMATION USING MACHINE LEARNING (ANN AND SVR) IN SMART BUILDINGS

Martin Gabriel[1], Thomas Auer[2]
[1] *Technische Universität München, Deutschland, E-Mail: martin.gabriel@tum.de*
[2] *Technische Universität München, Deutschland, E-Mail: thomas.auer@tum.de*

Abstract

This research reports on a case study in an open office, recording indoor air pollution and smart building data in high spatiotemporal resolution for time intervals amounting to more than six months and spread over four seasons. Six measurement nodes recorded indoor air pollution (particulate matter concentration) and smart-building data (temperature, pressure, humidity, sound, illumination, window opening states, and printer power consumption) on a sub-minute time resolution. The data was used to train machine learning models and to evaluate the predictions. Two machine learning typologies are here examined: artificial neural networks (ANNs) and support vector machine regression (SVRs). The models are tested for three individual weeks and evaluated using the R2 and mean average error (MAE) metric. ANNs were found to perform best with maximum R2 of 0.72 and a mean R2 of 0.68 accounting for all weeks.

Introduction

With over 90% of the time spent indoors in western countries, indoor air pollution poses a risk to health and psychological well-being (Sundell, 2004). WHO indicated that air pollution is a globally significant health threat and the foremost environmental factor for premature deaths in Europe (Henschel, Chan et al., 2013). According to the European Environment Agency exposure to fine particulate matter is responsible for more than 400,000 early deaths in European countries (Ortiz et al., 2019).

Particulate matter is a mixture of organic and inorganic compounds. It can be differentiated into coarse particulate matter (PM10) with an aerodynamic diameter above 2.5µm and fine particulate matter (PM2.5) with an aerodynamic diameter below 2.5µm (Maroni et.al., 1995). Some components of particulate matter have been found to cause cancer and thus declared as carcinogenic by the International Agency for Research on cancer (Loomis et al., 2013).

Therefore, legislators employ rules and guidelines regarding exposure times and concentration limits to secure occupant health in indoor environments. An in-detail summary of current guidelines and regulations in different countries is given in (Abdul-Wahab et.al., 2015). The WHO guidelines recommend a continuous exposure limited to 200 µg/m³ for PM10 and 10 µg/m³ for PM2.5 (Abdul-Wahab et.al., 2015).

Studies (Li et al., 2018, Szigeti et al., 2017) have found that within buildings and rooms concentration significantly differs based on source location, air change rate and air flow patterns. Compared to gaseous pollutants, this is especially pronounced for solid pollutants due to their lower propagation velocity as well as deposition and resuspension on building materials (Szigeti et al., 2017).

Appropriately assessing the risk to occupants by particulate matter, therefore, requires measurements with a high spatiotemporal resolution. However, high cost and sensitive optical measurement equipment makes ubiquitous deployment of sensors throughout the building stock infeasible. Therefore, alternatives to ubiquitous sensing must be found to assess particulate matter concentration.

Research gap

Monitoring indoor pollutant concentration is crucial for creating healthy office environments. It has been found in the literature that indoor particulate matter concentration varies significantly on a room-by-room as well as on a sub-room basis with significant temporal fluctuations. Therefore, spatiotemporal high-resolution monitoring is necessary to achieve an accurate occupant risk assessment. However, performing large scale measurements is inapplicable for widespread application due to the sensitivity and high cost of sensor equipment. Our research therefore focuses on estimating indoor pollutant concentration based on available smart building data on a sub room level using ML methods.

Measurement of indoor particulate matter

The predominant sources of indoor particulate matter are both external, as traffic and smog, and internal, as furniture, devices, and human effluents.

Optical particle counters (OPC) are common for monitoring particle pollutants (Manikonda et.al., 2016). Light of the same wavelength as the particle size is used to estimate particular matter concentration by the amount of scattered light. Due to its corresponding wavelength, infrared light is used for particles between one µm and ten µm. Particles significantly larger than the wavelength tend to absorb the light and particles significantly smaller do not interact, thus leading to a high specificity of the approach.

Machine learning for indoor pollution estimation

An alternative to sensor deployment is using machine learning (ML) for indoor pollution estimation (Ma et.al., 2021). ML is applicable if complex or indistinguishable relationships between input and output data impede the use of regular algorithms. Prominent fields of application include speech and image recognition as well as natural language processing.

To train the model, extensive data is required. For building stock, the necessary data is available in smart buildings. Smart building data is now the norm in new buildings and increasingly available in old buildings through retrofitting of wireless BUS systems. Smart buildings gather various kinds of data from actuators, the water, air and electric systems, indoor environment, and occupant activity.

Existing approaches for ML usage for smart buildings include prediction of occupancy, recognition of activity and optimization of energy efficiency by device control (Qolomany et al., 2019). As well as thermal comfort estimation and energy demand prediction (Djenouri et al., 2019).

Common supervised ML methods for smart building application include artificial neural networks (ANN) and support vector machines (SVM). SVMs transform data into multidimensional space, enabling the application of linear techniques to nonlinear problems (Karatzoglou et.al., 2006). ANNs imitate the working principle of neurons in the human brain. Several layers of neurons are connected to each other with variable weights and triggered by activation functions (Qolomany et al., 2019).

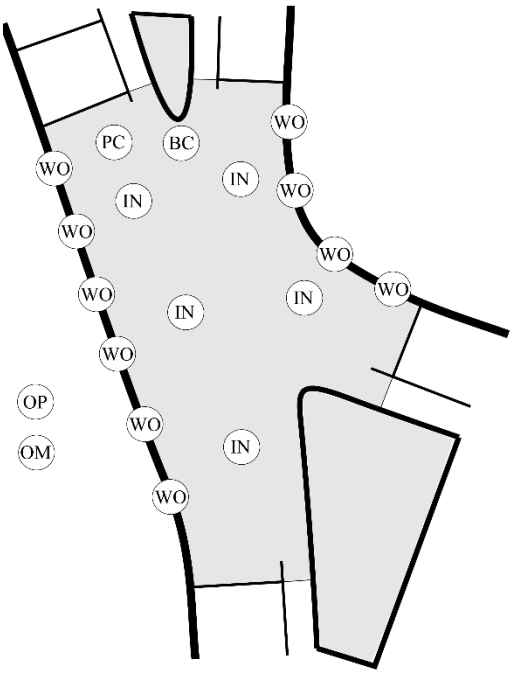
ML implementation for smart buildings is typically performed in four steps: data acquisition, data preprocessing, learning and interpretation. An in-depth description of the work steps is given in (Djenouri et al., 2019).

Methods

In this study two ML methods, ANNs and support vector machine regression (SVR) were explored for different smart building data measurement intervals. Ground truth and training data were collected with measurement nodes, from the building control system and local meteorological stations, in high spatiotemporal resolution and combined with contextual data from the HVAC-system. To account for different resolutions of smart building data, a high-resolution (10 second sampling) and a low-resolution (hourly sampling) scenario were compared and implemented for two corresponding ML methods: ANN, especially applicable for substantial amounts of data, and SVR, able to achieve predictions with smaller amounts of data.

Measurement setup

Measurements were performed in a typical open-plan-office in a high-rise building in Munich. The office is conditioned by thermally activated floors and ceilings, fresh air is introduced through ground level induction units and extracted at ceiling level. Air is supplied at a constant rate of 1.6 air changes per hour between 5.15 a.m. and 08.00 p.m.; additionally, operable windows can be used for fresh air supply. The open-plan office is located on the third floor and its two external façades are oriented towards north-west and south-east.



WO	Window state nodes	**IN**	Indoor node
OP	Outdoor pollution nodes	**BC**	Building control node
OM	Outdoor meteoro-logical nodes	**PC**	Power node

Figure 1: Measurement setup

Measurements were performed from June 2021 to December 2021. Measurements were collected with high spatiotemporal resolution, and with six measurement nodes located throughout the office sampling data on a sub-minute time resolution. The measurement nodes were placed according to the standardized EPA protocol to measure indoor air quality in large office buildings (EPA, 2003) at a distance greater than 0.5 meters from walls, corners, and windows and one meter from local sources and occupants, and not in front of or below air supply units and not in direct sunlight. All nodes were located at a height of 1.10 meters (breathing zone).

The measurement nodes collected indoor air pollutant data, smart building data was collected from the building control system and complemented by the measurement nodes in higher resolution. In total nine smart building data points (temperature, humidity, ambient air pressure, power consumption, window state, illumination, noise level) where collected.

The Sensirion SPS30 sensor is used for indoor pollutant measurements. The selection is based on the assessment of pollutant sensors in (Demanega et al., 2021). The study compared the performance of particulate matter sensors under different pollution scenarios in comparison to a reference instrument (miniWRAS). (Demanega et al., 2021) ascertained that the SPS30 achieved a pearson correlation coefficient above 0.8 for all pollution scenarios and, therefore, "very strongly correlated with miniWRAS" (Demanega et al., 2021). However, the sensor underreported particulate matter concentrations in some scenarios up to 73%.

Sensor calibration

To reduce sensor bias, sensors were cross calibrated. All sensors were placed close to each other in an indoor environment and gathered data over the course of 24 hours. A mean value for each measurement was calculated and a polynomial function produced using a regression for each sensor and measurement. The polynomial function was used to calibrate the raw measurement values.

Due to the SPS30s underreporting of particulate matter concentration, a calibration function was derived from the measurment in (Demanega et al., 2021) comparing SPS30 and miniWRAS data. A regression analysis was used to determine a linear function between SPS30 and miniWRAS measurements. The derived linear function was then applied to all SPS30 raw measurements in this study.

Data preprocessing

To achieve optimal prediction results, raw sensor data were preprocessed. The preprocessing procedure performed an enrichment of the data, handling of missing data, smoothing of fluctuations, and finally splitting the data in a test and training set.

Raw measurement data was complemented by adding contextual information for each timestamp, regarding HVAC operation schedule (5:15am to 8pm during workdays), a workday, weekend, and holiday classification as well as a tag indicating the season. Furthermore, open accessible weather (distance < 4km) and outdoor pollution (< 500m) data were integrated.

Initial tests showed a transient sensor response after power cycling, resulting in erroneous sensor readings. Therefore, measurements up to 15 minutes after a power cycle were cropped from the dataset. Power outage or connection loss led to brief interruptions of sensor node readings throughout the whole data set. This was mitigated by forward broadcasting the last valid measurement for up to ten missing timesteps. Due to varying sampling intervals of the sensors, raw data time intervals were variable. To ensure efficient processing, all sensor readings were resampled to a fixed interval of 10 seconds, by taking the mean value for each time-step.

Sensors are prone to fluctuations, due to measurement inaccuracies and minor environmental effects. To prevent misinterpretation in the machine learning model, random fluctuations were smoothed. Smoothing was implemented with an exponential weighted mean function, considering previous measurements with an exponential decaying weight and therefore, emphasizing trends over random fluctuations.

Finally, the dataset was split in a training and test set, used for training the machine learning model and evaluating its accuracy. Three working weeks (25.10 – 29.10; 8.11 – 12.11; 20.12 – 24.12) with at least four months of previous sampling and a typical occupancy profile (avoiding holiday periods) were selected for testing.

Feature selection and model training

It has been shown that the performance of machine learning models deteriorates if too many unimportant features are used in its training process (Djenouri et al., 2019). Therefore, it is crucial to preselect input features. In this study feature selection was implemented as a filtering approach using Pearson (linear) and Spearman (non-linear) correlation coefficients.

Correlation metrics were calculated between each potential feature and the output target (particulate matter concentration). Each feature with a correlation value higher than 0.2 for either Pearson-correlation coefficient or Spearman-correlation coefficient were selected as input value for the machine learning process. Even though a preselection regarding correlation coefficients was performed, further improving the machine learning model by reducing input data dimensionality has been suggested in the literature (Djenouri et al., 2019).

In this study principal component analysis (PCA) was employed to reduce input data dimensionality whilst explaining 95% of its variance by reducing observation redundancy and noise.

Previous studies have tended to deploy two predominant machine learning methods for smart building applications: artificial neural networks and support vector machines (Djenouri et al., 2019). Both were implemented in this study and their respective outputs compared to each other in the results section.

Model evaluation

The trained machine learning model was evaluated using the test dataset over three weeks. Its predictions were compared to the measured ground truth and evaluated with selected prediction metrics. To

evaluate the fit of the prediction algorithm the R2 metric (coefficient of determination) was used, and the MAE metric for evaluation of the mean error.

The R2 metric indicates the variance of the predicted results compared to the truth values. Its optimum of 1.0 indicated a perfect fit. Equation 1 shows the calculation of the R2 metric.

The MAE (mean absolute error) indicates the mean error between prediction and truth. Its scale is therefore dependent on the range of values of the ground truth data, the lower MAE the better the prediction. Equation 2 shows the calculation of the MAE metric. The R2 and MAE metrics were calculated once for each individual testing week as well as aggregated for all testing weeks.

$$R2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y_i})^2}{\sum_{i=1}^{N}(y_i - \overline{y}_i)^2} \quad \textbf{(1)}$$

$$MAE(y,\hat{y}) = \frac{\sum_{i=0}^{N-1}|y_i - \hat{y_i}|}{N} \quad \textbf{(2)}$$

y_i = measured value; $\hat{y}$ = predicted value; $\overline{y}$ = mean of measured values; N = number of samples

Results

In the following section, prediction results for the three test weeks are shown differentiated for ANN and SVR as well as high resolution and low-resolution data. The figures show the results of each week as well as their corresponding accuracy metrics. The red line in the plots presents the measured truth, the blue lines are the predicted values as given by the machine learning model. Table 1 summarizes the prediction accuracy metrics.

Overall, model processing and training was performed for 1,248,975-time steps for high resolution and 4,848-time steps for low resolution on an 8-core machine with 20GB of memory, requiring 259 and 16 seconds, respectively, for high resolution and low-resolution data of ANN, 522 and 20 seconds for SVR. Indicating feasibility for multi-room, multi-building deployment.

In summary the highest correlation with the lowest error was achieved using ANNs. Low-resolution and high-resolution observation data performed similarly for ANNs with a slightly better performance for low resolution data. SVRs performed almost as well as ANNs in the case of high-resolution data in terms of the R2 metric, though at a higher MAE. For Low resolution data, SVRs performed significantly worse.

Looking at the variance of results between the individual weeks, ANNs performed more dependably while exhibiting the lowest accuracy in the second week of testing. SVRs showed higher fluctuations in their results, also performing worst in the second week.

The maximum prediction accuracy of 0.72 R2 was achieved in the third testing week for ANN with low resolution data, minimal error for SVR with an MAE of 0.79. To summarize, ANNs performed better and more dependably under all scenarios, being able to handle different observation resolutions.

Transferability

Being able to transfer the trained model to other buildings is key to minimize setup effort and sensor deployment. Due to the multitude of factors affecting indoor pollutants as building location, construction materials, and occupant activity, it is necessary to allow for a tuning phase to fit the model to the new environment, providing measured pollutant data.

However, since the underlying processes are already ingrained, the tuning phase requires significantly less data than the initial training phase and allows for a temporary installment of sensors in contrast to permanently deploying them.

Furthermore, the integration of static building parameters as geometry, occupation density, orientation and used materials as well as the continuous improvement of the model using the additional measurements is expected to further reduce the tuning phase and improve accuracy.

Table 1: Prediction accuracy

		W1		W2		W3		All	
		ANN	SVR	ANN	SVR	ANN	SVR	ANN	SVR
HighRes	R2	0.66	0.69	0.55	0.39	0.56	0.64	0.64	0.62
	MAE	1.53	1.74	1.19	1.55	1.14	1.14	1.29	1.48
LowRes	R2	0.70	0.56	0.52	0.46	0.72	0.62	0.68	0.57
	MAE	1.55	1.62	1.29	1.34	1.27	1.02	1.25	1.32

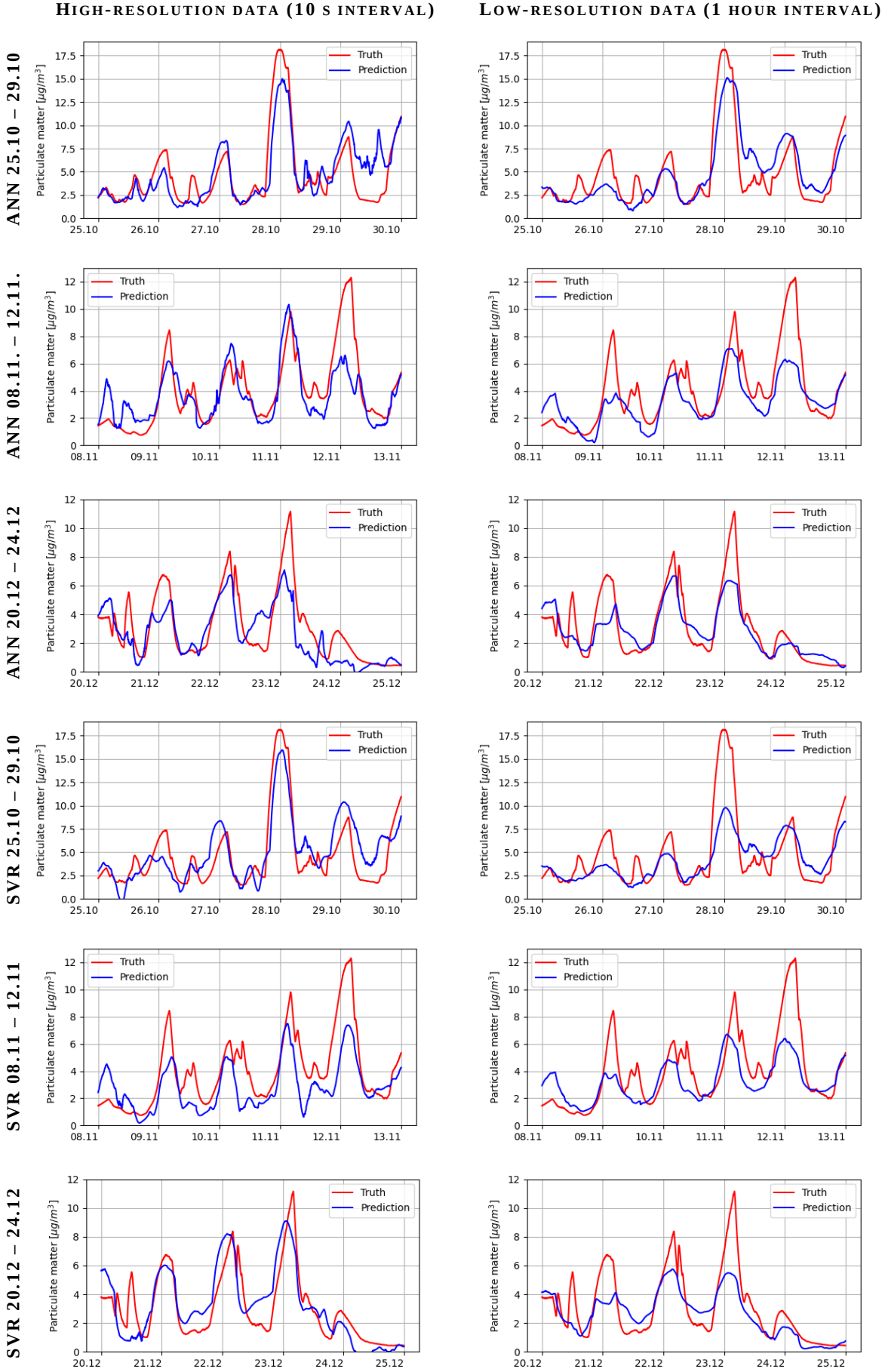


Figure 2: Prediction results differentiated by testing week, model, and resolution

Conclusion

Spatiotemporal high resolution indoor pollutant monitoring is crucial for creating healthy office environments since literature indicates significant spatiotemporal variations within buildings and rooms. However, the sensitivity and high cost of sensor equipment impede the wide-scale application of measurement devices. The aim of our research was to estimate indoor pollutant concentration based on available smart building data on a sub room level using ML methods.

An ML pipeline was developed for ANN and SVR, utilizing six months of measurement data. The result of a mean R2 of **0.64** and **0.68** indicates a strong correlation of the prediction models with the ground truth and MAEs of **1.29** and **1.25** show that the mean error between the predicted model and ground truth only constituted 10% of the ground truth variance for the better performing ANNs. Therefore, we found that commonly available smart building data can be used to predict indoor pollutants to a high degree of accuracy, thus rendering spatiotemporal high-resolution measurements unnecessary in monitoring and risk assessment for occupants. Furthermore, prediction models can be integrated in the control of demand-controlled ventilation or provide users guidance whether to open operable-windows in naturally ventilated buildings, thus enabling not only monitoring but active mitigation without the need for retrofitting measurement equipment. Therefore, our prediction model supports improving and monitoring the indoor air quality in naturally as well as in mechanically ventilated buildings.

Further research will focus on expanding the methodology on further nonresidential typologies and testing the model's transferability. Furthermore, relevant indoor air pollutants as CO2 and VOC should be implemented in the methodology.

References

Abdul-Wahab, Elkamel, A., Ahmadi, L., and Yetilmezsoy, 2015. A review of standards and guidelines set by international bodies for the parameters of indoor air quality, Atmospheric Pollution Research, Elsevier

Demanega, I., Mujan, I., Singer, B. C., Andelkovic, A. S., Babich, F., and Licina, 2021. Performance assessment of low-cost environmental monitors and single sensors under variable indoor air quality and thermal conditions, Building and Environment, (187), Elsevier

Djenouri, D., Laidi, R., Djenouri, Y., and Balasingham, 2019. Machine learning for smart building applications: Review and taxonomy, No. 2, ACM Computing Surveys (CSUR), ACM New York, USA

EPA, A. 2003. A standardized EPA protocol for characterizing indoor air quality in large office buildings, Indoor Environment Division US EPA, Washington, DC, USA.

Karatzoglou, A., Meyer, D., and Hornik, K. 2006. Support vector machines in R, Journal of statistical software.

Li, J., Li, H., Ma, Y., Wang, Y., Abokifa, A. A., Lu, C., and Biswas, P., 2018. Spatiotemporal distribution of indoor particulate matter concentration with a low-cost sensor network, Building and Environment, Elsevier.

Loomis, D., Grosse, Y., Lauby-Secretan, B., El Ghissassi, F., Bouvard, V., Benbrahim-Tallaa, L., Guha, N., Baan, R., Mattock, H., and Straif, K., 2013. The carcinogenicity of outdoor air pollution, Lancet Oncology, Elsevier Limited.

Ma, N., Aviv, D., Guo, H., and Braham, 2021. Measuring the right factors: A review of variables and models for thermal comfort and indoor air quality, Renewable and Sustainable Energy Reviews, Elsevier.

Manikonda, A., Zíková, N., Hopke, P. K., and Ferro, A. R. 2016. Laboratory assessment of low-cost PM monitors, Journal of Aerosol Science, Elsevier.

Maroni, M., Seifert, B., and Lindvall, T. 1995. Indoor air quality: a comprehensive reference book, Elsevier.

Ortiz, A. G., Guerreiro, C., Soares, J., Antognazza, F., Gsella, A., Houssiau, M., 2019. Air Quality in Europe—2019 Report, European Environment Agency, Luxembourg.

Qolomany, B., Al-Fuqaha, A., Gupta, A., Benhaddou, D., Alwajidi, S., Qadir, J., and Fong, 2019. Leveraging machine learning and big data for smart buildings: A comprehensive survey, IEEE.

Sundell, J. 2004. On the history of indoor air quality and health, Indoor air.

Szigeti, T., Dunster, C., Cattaneo, A., Spinazzè, A., Mandin, C., Le Ponner, E., de Oliveira Fernandes, E., Ventura, G., Saraga, D. E., Sakellaris, I. A., 2017. Spatial and temporal variation of particulate matter characteristics within office buildings—The OFFICAIR study, Science of The Total Environment, Elsevier.