

Optimisation of thermal comfort and indoor air quality estimations applied to in-use buildings combining NSGA-III and XGBoost

Miguel Martínez-Comesaña^{a,*}, Pablo Eguía-Oller^a, Javier Martínez-Torres^b,
Lara Febrero-Garrido^c, Enrique Granada-Álvarez^a

^a Department of Mechanical Engineering, Heat Engines and Fluids Mechanics, Industrial Engineering School, University of Vigo (Universidade de Vigo), Maxwell s/n, 36310 Vigo, Spain

^b Department of Applied Mathematics I, Telecommunications Engineering School, University of Vigo (Universidade de Vigo), 36310 Vigo, Spain

^c Defense University Centre, Spanish Naval Academy, Plaza de España, s/n, 36920 Marín, Spain

ARTICLE INFO

Keywords:

IEQ
Air quality
Thermal comfort
XGBoost
NSGA-III

ABSTRACT

Indoor environmental quality (IEQ) monitoring of in-use buildings has become essential in recent years due to the COVID-19 pandemic, as it significantly affects the well-being, health and productivity of building users. Nevertheless, knowing in real time the environmental conditions in large multi-zone areas is a difficult issue. Thus, the use of machine learning techniques to estimate indoor conditions has increased considerably. The aim of this paper is to present an interpolation model, based on an optimised extreme gradient boosting algorithm, to estimate every minute the indoor temperature, relative humidity and CO₂ concentration inside buildings. These estimations are obtained without requiring permanent monitoring in the occupied zone. The optimisation, focused on finding the minimum number of monitoring devices needed to provide accurate interpolations, is performed using the multi-objective genetic algorithm NSGA-III. This methodology was applied in a research centre in the north-western Spain. The results show that the optimised or reduced model is capable of estimating indoor temperatures and relative humidity with relative errors below 6% and CO₂ levels below 10%.

1. Introduction

Meeting users' needs is an important issue in the Architecture, Engineering, Construction and Operation (AECO) field (Roberts, Edwards, Hosseini, Mateo-Garcia, & Owusu-Manu, 2019). Taking into account that people stay 80%–90% of the time indoors (ASHRAE, 2011), indoor environmental quality (IEQ) has become an important factor in numerous research studies in past years (Costantino, Fabrizio, Ghiggini, & Bariani, 2018; Martínez-Mariño, Eguía-Oller, Granada-Álvarez, & Erkoreka-González, 2021; Yu, Song, Song, & Liu, 2021). An adequate indoor environment depends on factors such as air quality or visual, thermal and acoustic comfort that strongly influence in health, productivity and well-being of building users (Alavi, Forcada, Bortolini, & Edwards, 2021; Andargie & Azar, 2019). On the one hand, these IEQ-based factors may be used to define an appropriate range of comfort levels (Wang, Yan, & Xiao, 2012). Indeed, the World Health Organisation (WHO) states that about 4.3 million people died prematurely each year due to indoor air pollution (Ritchie & Roser, 2013). Thus, IEQ has been incorporated into localised green building certification systems such as LEED and BREEAM in recent years. (Awadh, 2017; Elnaklah,

Walker, & Natarajan, 2021). On the other hand, IEQ and energy consumption in buildings are related since the maintenance of acceptable indoor conditions depends on cooling, heating or ventilation systems. Thus, controlling IEQ is essential to analyse whether the decrease in the energy consumption is negative for comfort conditions (Steinemann, Wargocki, & Rismanchi, 2017).

Considering the current situation where, due to the COVID-19 pandemic, air quality monitoring becomes a key factor and thermal comfort is considered the most important IEQ category by building users (Frontczak & Wargocki, 2011), this paper focuses on indoor air quality and indoor thermal comfort. In recent years, different studies have attempted to analyse the influence of alternative indoor environmental conditions on these two IEQ categories (Andargie & Azar, 2019; Ma, Aviv, Guo, & Braham, 2021; Park, Loftness, Aziz, & Wang, 2019; Sant'Anna, Dos Santos, Vianna, & Romero, 2018). Specifically, the aforementioned studies conclude that measuring indoor concentration of pollutants (such as CO₂) is an efficient way to monitor air quality and that both indoor air temperature and relative humidity are the most frequently used variables to assess thermal comfort. Moreover,

* Corresponding author.

E-mail address: migmartinez@uvigo.es (M. Martínez-Comesaña).

Abbreviations

IEQ	Indoor Environmental Quality
XGBoost	Extreme Gradient Boosting
MLP	Multi-Layer Perceptron neural network
GA	Genetic Algorithm
NSGA	Nondominant Sorting Genetic Algorithm
CV(RMSE)	Coefficient of Variation of the Root Mean Squared Error
NMBE	Normalised Mean Bias Error
ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers
CINTECX	Centre for Research in Technology, Energy and Industrial Processes
HVAC	Heating, Ventilating and Air Conditioned
GFS	Global Forecast System

several previous studies were based on these variables to analyse the thermal comfort and air quality inside a building (Hesaraki, Myhren, & Holmberg, 2015; Martínez-Comesaña et al., 2021; Shrestha, Rijal, Kayo, & Shukuya, 2022; Yu et al., 2021). Knowing and monitoring these variables enables improving the health and comfort of the occupants on the one hand, and the energy efficiency of the building on the other hand. The most widely used method for IEQ control is the installation of monitoring devices at fixed locations (Guerra-Santin & Tweed, 2015; Vilcekova et al., 2017; Zuhaib et al., 2018). These methods show problems in large areas, where the indoor environment is not as uniform as in small areas. Therefore, the use of mobile devices for collecting measurements at different positions has been extended (Jin et al., 2017; Yang, Liu, Wang, Cao, & Li, 2021).

An accurate knowledge of the IEQ of a building is an important first step towards improving its environmental quality. In this regard, different approaches have been explored in the past few years to tackle this problem. One of them is the improvement of indoor monitoring systems, based on more precise sensors and devices or less interfering positioning, to know more accurately the evolution of the environment inside buildings (Han, Gao, & Fan, 2012; Wu, Hou, Shen, & Lian, 2021). Furthermore, models based on spatial visualisation (Yu et al., 2021), statistics (Shen et al., 2021) or building simulations (Ganesh, Sinha, & Verma, 2020; Gilani, Montazeri, & Blocken, 2016; Martínez-Mariño et al., 2021) (known as white-box models) have also been used to try to precisely estimate the IEQ of a building. In addition, the use of complex mathematical models such as machine learning models (known as black-box models) has expanded significantly in this context. These type of models stands out from the rest due to their robustness, strong yield, ease of implementation and the fact that specific knowledge of the subject is not required to extract complex patterns of the data (Gong et al., 2020; Martínez Comesaña, Febrero-Garrido, Troncoso-Pastoriza, & Martínez-Torres, 2020; Martínez-Comesaña et al., 2021; Shan & Yang, 2020). Among these black-box models, extreme gradient boosting (XGBoost) is recognised for its efficiency and flexibility. This algorithm is an advanced version of the classic gradient boosting with more computational power and aiming to better avoid overfitting problems (Kumari & Toshniwal, 2021). XGBoost has already been applied in numerous scientific fields such as economics (Carmona, Climent, & Momparler, 2019; Chen, Li and Sun, 2020), biology (Babajide Mustapha & Saeed, 2016), building energy efficiency (Lu, Cheng, Ma, & Hu, 2020; Wang, Lu and Li, 2019) or environmental research (Jiao et al., 2020; Kumari & Toshniwal, 2021; Yu et al., 2020). Moreover, the use of genetic algorithms (GA) to optimise these models has recently increased (Chen, Zhao, Sun and Yin, 2020; Martínez-Comesaña et al., 2021; Sang, 2021). Specifically, the Non-Dominant Sorting Genetic Algorithm (NSGA-III) is an efficient

multi-objective algorithm, created to solve optimisation problems with more than two objectives, which allows focusing on the accuracy of several models at the same time or even on other variables related with the complexity of the model (Li, Deb, Zhang, Suganthan and Chen, 2019; Yin, Zhao, Lin, & Karcianas, 2020; Zhu, Liang, Chen, & Ming, 2017).

The aim of this paper is to present an optimised methodology to spatially and temporally estimate the IEQ of an in-use building combining the NSGA-III and the XGBoost algorithm. The study is conducted considering indoor air temperature, relative humidity and CO₂ concentration, but this methodology is applicable to any other variable related to IEQ. The monitoring installation is composed by sensors installed in 8 wall-mounted devices and in a mobile device. Additionally, weather variables of the building area together with certain temporal variables are also considered as model inputs. The optimisation using NSGA-III, focused on feature selection and adjustment of the XGBoost parameters, is based on minimising four objectives: the Coefficient of Variation of the Root Mean Squared Error (CV(RMSE)) obtained for the three variables of interest and the number of devices used to interpolate the data. Thus, a comparison between the original model and the optimised model is carried out. The metrics selected to compare the performance of each of the built models are the computational time required in the training process, the Normalised Mean Biased Error (NMBE) and the CV(RMSE) in accordance with the recommendations of the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE).

Previous studies have presented the use of XGBoost to estimate indoor environmental conditions (Huang, Miles, & Zhang, 2020; Saini, Dutta, & Marques, 2020) or the use of genetic algorithms to optimise the monitoring system or hyperparameters related to machine learning models (Komasi & Goudarzi, 2020; Martínez-Comesaña et al., 2021). The novelty of this paper lies in the implementation of XGBoost combined with NSGA-III, which simultaneously optimises the internal structure of the algorithm and the number of monitoring devices used to generate spatiotemporal estimations of the IEQ in a large multi-zone open space. This research contributes with a methodology that efficiently reduce the need of data collection and provides a real-time image of the distribution of the studied variables across the analysed space. Therefore, only a few collection points based on low-cost sensors are needed, which can be placed outside the occupation area, reducing significantly the monitoring cost. Therefore, the presented method becomes a crucial element that enables a proper evaluation of the indoor environment conditions of the analysed building and the information that it provides could be integrated in its digital twin representation.

2. Material and methods

This study is focused on the optimisation of thermal comfort and indoor air quality estimations based on the number of monitoring devices and specific hyperparameters of the machine learning model used. In a previous study (Martínez-Comesaña et al., 2021) the authors have shown that, considering an unoptimised installation with all available devices, accurate estimates can be obtained. However, as an independent study, this paper aims to improve this original research by different means as presented in Table 1.

Concretely, the main objective of this research is to present a methodology for optimising an existing monitoring installation by reducing the number of collecting devices required to be installed. At the same time, the optimisation is also focused on tuning the hyperparameters of the machine learning model, used to interpolate data, considering the variation in the number of devices providing data.

Table 1

Comparison between the original research conducted in the first part of 2021 and an advanced research conducted in the last part of 2021.

	Original research	Advanced research	Additional information
Objective	Generate accurate interpolations.	Optimising a monitoring system.	While in the original study the aim was to demonstrate the usefulness of the monitored data together with a machine learning model that interpolates data, in the new study the aim is to optimise the monitoring system as well as to improve the accuracy of the estimations.
Optimisation	MLP architecture	Number of devices together with certain hyperparameters of the XGBoost model.	Optimisation in the new research focuses on reducing the monitoring required. This is significant in reducing costs.
Monitoring period	18/10/2020–10/03/2021	18/10/2020–06/05/2021	As the system continued working, the available monitoring period for training the model is longer.
Algorithm	MLP	XGBoost	Intermediate studies have concluded that an algorithm based on decision trees is more efficient.
Genetic algorithm	NSGA-II	NSGA-III	NSGA-III is more efficient at optimising more than two objectives simultaneously.

2.1. Extreme gradient boosting

Extreme gradient boosting (XGBoost) is a sophisticated supervised algorithm widely recognised for its efficiency and flexibility. It can be implemented in a parallel and distributed way, enabling a faster learning process. The loss function related to XGBoost provides an additional regularisation term that contributes to smoothing the final learning weights and reducing the probability of overfitting. It also considers gradients up to the second order to optimise the loss function. Additionally, in order to avoid overfitting problems, XGBoost also handles row and column sampling (Gao et al., 2017). The following paragraphs explain how this algorithm works (Babajide Mustapha & Saeed, 2016; Yu et al., 2020; Zhang, Wu, Zhong, Li, & Wang, 2021).

This algorithm is based on an ensemble of K regression trees $T_1(\mathbf{x}_i, y_i), \dots, T_K(\mathbf{x}_i, y_i)$ where $\mathbf{x}_i$ is the training set of features associated with the sample i and y_i represents the real value to predict. The estimated output of the model (Eq. (1)) is obtained as the sum of the individual prediction outcomes ($f_k(\mathbf{x}_i)$) of all built trees:

$$\hat{y}_i = \sum_{k=1}^K f_k(\mathbf{x}_i), \quad f_k \in T \quad (1)$$

where T is the space of all regression trees and K is the number of regression trees. For a certain dataset, a prediction outcome is generated $f_k(\mathbf{x}_i)$ (known as *leaf score* or *leaf weight*) for each leaf node z . This leaf weight ω_z corresponds with the regression value of all samples at the leaf node z in the tree, where $z \in \{1, 2, 3, \dots, Z\}$ and Z represents the number of leaves in the tree.

The regularised objective function to optimise is defined in Eq. (2):

$$\Phi = \sum_{i=1}^n l(y_i, \hat{y}_i) + \gamma Z + \frac{1}{2} \lambda \sum_{z=1}^Z \omega_z^2 \quad (2)$$

where $\sum_{i=1}^n l(y_i, \hat{y}_i)$ is the training loss function, n the number of samples and $\gamma L + \frac{1}{2} \lambda \sum_{z=1}^Z \omega_z^2$ is the regularisation term which penalises the complexity of the model. In this way, λ is the regularisation hyperparameter, γ is the complexity cost of adding a new leaf and ω_z^2 is the L2 norm of the weights in the leaf node z .

In an additive learning processes, the trees are built sequentially. The new trees learn and update the prediction residues from the previous ones (Martínez-Comesaña, Febrero-Garrido, Granada-Álvarez, Martínez-Torres, & Martínez-Mariño, 2020). Thus, in the k th iteration, $\hat{y}_i^k$ is calculated as $\hat{y}_i^{k-1} + f_k(\mathbf{x}_i)$. Thus, after several operations described in Babajide Mustapha and Saeed (2016), Chen and Guestrin (2016) and Zhang et al. (2021) and considering a fixed tree structure, the optimal leaf weight ω_z^* and the optimal objective function Φ_k^* , which indicates how good the model structure is, are given by Eqs. (3) and (4), respectively:

$$\omega_z^* = - \frac{\sum_{i \in I_z} g_i}{\sum_{i \in I_z} h_i + \lambda} \quad (3)$$

$$\Phi_k^* = - \frac{1}{2} \sum_{z=1}^Z \frac{\left(\sum_{i \in I_z} g_i \right)^2}{\sum_{i \in I_z} h_i + \lambda} + \gamma Z \quad (4)$$

Further information about XGBoost is available in Chen and Guestrin (2016) and Thongsuwan, Jaiyen, Padcharoen, and Agarwal (2021).

2.2. Non-dominated sorting genetic algorithm III

This genetic algorithm is an improvement, based on the selection mechanism, of NSGA-II (Campos-Ciro, Dugardin, Yalaoui, & Kelly, 2016; Martínez-Comesaña et al., 2021). The main procedure of the NSGA-III starts by defining some concrete reference points. An initial population (P_t) of size N is then randomly extracted and an iteration process begins and continues until some stopping criterion is reached. This iteration process is described below (Cui, Chang, Zhang, Cai, & Zhang, 2019; Parvizi & Rezvani, 2020; Yi et al., 2018):

1. Considering the iteration r th, an offspring population O_t is generated from the population P_t . This is done through a simulated binary crossover (SBX) (Jiang et al., 2016), a random selection (Zhu et al., 2017) and a polynomial mutation (Li, Sheng, Zhi and Li, 2019). The size for the two populations stays in N .
2. P_t and O_t are merged in a new population $M_t = P_t \cup O_t$ with size $2N$. First, the best N members are selected for the next generation based on a non-dominated sorting that classifies the elements of M_t in different non-domination levels (L_z).
3. A new population G_t is formed by adding the elements of M_t , level by level, until the size of N is reached or exceeded for the first time. Assuming that the last level considered is the level z , the levels from $z+1$ to the last one created are dismissed.
4. If the size of G_t , considering the last selected levels, is exactly N all the elements are chosen for P_{t+1} . If the size of G_t including the level z is greater than N , elements in $G_t \setminus L_z$ are chosen directly for P_{t+1} and the gaps until N are selected from L_z maintaining the diversity in the new population.

In comparison with NSGA-II, the selection process with NSGA-III change to a more complex analysis of the members in L_z in relation to the supplied reference points. In addition, a normalisation of the objective values and the reference points, so they have the same range, is required. Afterwards, the ideal point considered to G_t is the zero vector. Therefore, the perpendicular distances between the elements of G_t and the reference lines, linking the ideal point with the reference points, are calculated (see Fig. 1). Each of the members in G_t is associated with the specific reference point that yields the smallest perpendicular distance. Thus, the operator α_j is defined as the number of elements of $G_t \setminus L_z$ related with the j th reference point (*niche count*).

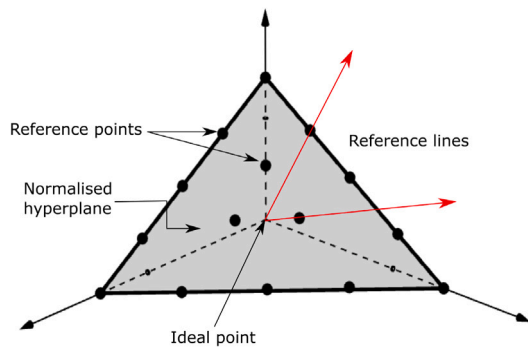


Fig. 1. NSGA-III selection process. Structure of reference points and reference lines at a normalised reference level.

5. The selection of the elements of L_z (*niche-preservation*) is carried out following the steps described in Cui et al. (2019).

In this way, this algorithm forms a Pareto front with the best factible points among all the options tested. The decision function used to select an unique optimal point among the points in the Pareto front is the *pseudo-weights* function due to the appropriability with this specific genetic multi-objective algorithm (Blank & Deb, 2020). Lastly, to study the convergence of the algorithm and, therefore, the reliability of the results, the *centre distance* indicator (Wang et al., 2019) is used. This indicator is widely used when the ideal solutions are not available and it has the following form:

$$CD = \frac{1}{N} \sum_{n=1}^N \sqrt{\sum_{obj} f_{n,obj}^2} \quad (5)$$

This indicator judges how the population values change between generations (N) and helps to measure whether the number of evaluations is enough observing the distance between the origin point and the centre of the Pareto front. Further information about NSGA-III and its implementation can be consulted in Parvizi and Rezvani (2020) and Yi et al. (2018).

2.3. Case study data acquisition

This section describes the studied building, the data collection and the data pre-processing needed to perform the analysis presented in the next section.

2.3.1. Description of the building

The building analysed in this article is the Research Center in Technologies, Energy and Industrial Processes (CINTECX) located at the University of Vigo in north-western Spain and north-west oriented (see Fig. 2). Several research groups from different areas such as energy, sustainability, electronics or automation, develop their activities in this building. The diversity of work makes it interesting for this study. Concretely, the study is focused on the ground floor, where the laboratories are situated and the measuring devices are installed. Although there is no specific schedule for workers and experiments, activities normally take place between 8 a.m. and 8 p.m. This room, which consists of several different working areas where the research groups conduct their experiments, is an open space of 825 m² and 6 m of height. Therefore, there is a need for different equipment such as heaters, boilers or welding equipment, which can have an impact on the comfort of the occupants and affect the air quality. This equipment, some of which is located in closed laboratories, is the main source of heat gains from the floor. The HVAC (Heating, Ventilating and Air Conditioned) system working in the analysed space comprises four fan coils placed across

Table 2

Technical specifications of sensors considered in this study.

Sensor	Variable	Measuring range	Accuracy	Resolution
SHT31-D	Temperature	−40 to 125 °C	±0.3 °C	0.015 °C
	Relative humidity	0 to 100%	±2%	0.01%
MHZ-14	CO ₂	0 to 5000 ppm	±50 ppm ±5%	10 ppm

the room (see Fig. 4). This floor has a 35.5 cm wide concrete envelope with numerous windows occupying approximately 40% of the entire envelope. Thus, natural light is significantly exploited in this space. Fig. 2 shows pictures of the outside and some interior areas of the building. More information about the building is available in a previous study (Martínez-Comesaña et al., 2021).

There are two sources from which the data required for this study is collected. On the one hand, 9 monitoring devices gather data inside the building. On the other hand, the nearby meteorological station, supplemented by a numerical weather prediction model in case data are unavailable, provides weather data outside the building.

2.3.2. Monitored data

The indoor monitoring system considered in this analysis consists of 9 custom-made devices: one mobile device and 8 wall-mounted devices. These devices contain sensors to monitor, on the one hand, indoor air temperature and relative humidity, as these are commonly monitored variables and are closely related to the thermal comfort of a building. On the other hand, the devices also have sensors to monitor CO₂ as it is an effective indicator of air quality and room ventilation and, in recent years, has become an essential parameter due to the COVID-19 pandemic (Martínez-Comesaña et al., 2021; Park et al., 2019). All the collecting devices have installed the same sensors. Temperature and relative humidity are collected by SHT31-D sensor whereas CO₂ levels are measured by MHZ-14 sensor. Taking into account that outdoor CO₂ data are not available and that the building is located in a non-urban area, we have calibrated the indoor CO₂ sensors based on the average value obtained from measurements, taken at different times, of the CO₂ concentration outside the building. Their technical details are presented in Table 2.

The mobile device (D0) is used to gather data from multiple points across the analysed room. The wall-mounted devices (D1 to D8) are placed in fixed locations (where users are not disturbed) at heights within the range of 1.9 and 3.2 m. This equipments are characterised by low-cost hardware components (the price of the most expensive sensor does not exceed 35 €) and low-cost manufacturing. They are mainly manufactured by a 3D printing using polylactic acid (PLA) (see Fig. 3).

The distribution of the wall-mounted devices through the building, together with the positions monitored by the mobile device are shown in Fig. 4. In this case, only areas where building users work are taking into account. Closed laboratories are not studied since they have specific environmental conditions inside and are usually not occupied during experiments.

2.3.3. Weather data

The meteorological data source in this study is an automatic weather station of MeteoGalicia weather agency. Its location is 300 m north-east from the centre of the analysed building, with an ascending height differential of 43 m. In particular, the weather variables gathered from this station are the average air temperature, relative humidity and global solar irradiance (measured at 1.5 m of height from local ground surface).

The data from the mentioned station are obtained using a RSS service supplied by MeteoGalicia (de Galicia, 2021). Their frequency is in 10-min intervals with a delay of 15–25 min from real time. In addition, for replacing invalid or missing values from the selected station, the Global Forecast System surface flux (GFS sflux) model is used as a backup source. GFS is an international numerical weather



Fig. 2. Photographs of the studied building.

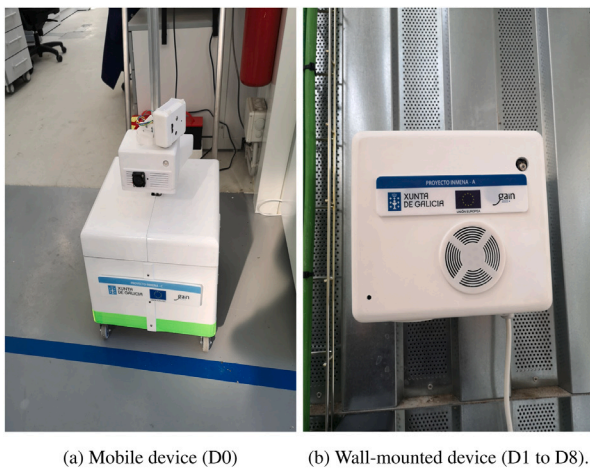


Fig. 3. Measuring devices used in this study.

prediction model built by the National Oceanic and Atmospheric Administration (NOAA) in USA (NOAA, 2021b). Specifically, GFS flux products obtain predictions each hour on a horizontal grid of 13 km resolution points. A more detailed information about GFS can be found in NOAA (2021a) and further information on this specific GFS application in Martínez-Comesaña et al. (2021).

2.3.4. Data preprocessing

This study is focused on the estimation of temperature, relative humidity and CO₂ values across a large multi-zone open space inside a building. The available data are minute observations of the three selected variables collected from the sensors installed in the wall-mounted devices and in the mobile device (see Table 2). The monitoring period went from November 18, 2020 to May 6, 2021; conforming a sample of approximately 240,000 observations (although in many cases not all sensors were collecting data) and considering 34 positions of the mobile device (see Fig. 4). The mobility of this device allows the introduction of its distances to fixed devices as model inputs. These distances are expressed in three dimensions (x-, y- and z-axes) as the mobile device also presents a vertical movement taking measurements at 5 different heights (55, 85, 115, 145, 175 cm). Thus, the black box models are trained by relating the data from wall-mounted devices, the mentioned distances, the meteorological conditions in the area and three time variables (day of the year, day of the week and hour of the day), with the values measured by the mobile device (desired values) (see Table 3). For CO₂ values (model 3) the available meteorological variables are not considered because the correlation between them and indoor CO₂ concentration is practically non-existent.

Other variables highly correlated with CO₂ measurements such as air exchange rate or occupancy rate are also not considered in this analysis. The justification for this is based on the greater difficulty for monitoring them (Franco & Leccese, 2020; Tien, Wei, Calautit, Darkwa, & Wood, 2020), similar previous studies which support this selection of model inputs (Costantino et al., 2018; Martínez-Comesaña et al., 2021; Yu et al., 2021), and, as presented below, the accuracy of the built models is comparable to that obtained in studies where the aforementioned variables are taken into account (Sharma et al., 2021).

Furthermore, the data collected by the mobile device has to be time-interpolated due to the fact that, during the vertical movement, it does not take measurements. With the aim of reducing the influence of the commented interpolation and sensor errors on the data, the devices data are smoothed applying the moving average technique (Babu & Reddy, 2014) (see Fig. 5).

Regarding the weather data, the collected observations are replicated each 10-min interval to obtain the data at the frequency required by the built models. Moreover, to take into account the internal inertia of buildings, a one-minute time lag is introduced as an ancillary input to the model (Martínez Comesaña et al., 2020).

This research focuses on the comparison between a set of original models and its optimal reduced versions, which structure is obtained with the NSGA-III. In this case, depending on the number of wall-mounted devices considered in the model (from 8 to 3) the number of non-operating wall-mounted devices supported by the models varies (see Table 4).

2.4. Error measurements

The error metrics under consideration in this study to assess the accuracy of the built XGBoost models are the Normalised Mean Bias Error (NMBE) and the Coefficient of Variation of the Root Mean Square Error (CV(RMSE)):

$$NMBE = 100 \times \frac{\sum_{i=1}^N (y_i - \hat{y}_i)}{\sum_{i=1}^N (y_i)}, \quad (6)$$

$$CV(RMSE) = 100 \times \frac{\sqrt{\sum_{i=1}^N (y_i - \hat{y}_i)^2 / N}}{\bar{y}}. \quad (7)$$

These validation metrics were already used in numerous similar studies (Martínez-Comesaña et al., 2020; Martínez Comesaña et al., 2020; Sha et al., 2019), which have shown the importance of low values for considering a model to be accurate. Moreover, in Martínez, Eguía, Granada, Moazami, and Hamdy (2020) is shown that the CV(RMSE) is the most appropriate measure in building simulation analysis (recommended by ASHRAE Ruiz & Bandera, 2017).

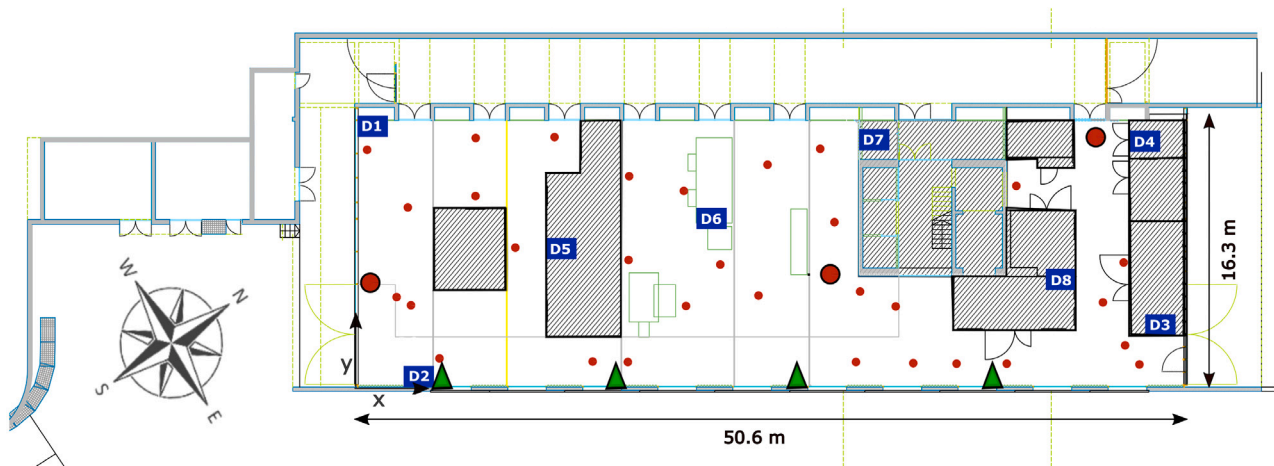


Fig. 4. Blueprint of the analysed area. The blue boxes indicate the positions of the fixed devices, the red dots represent the positions where the mobile device was placed, the green triangles indicate the fan coils positions and the hatched grey areas correspond to closed laboratories that are not considered in this analysis. Bigger dots represents the positions considered in the optimisation problem. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3

Inputs and outputs structure of the built models. The different type of variables introduced are also shown. The outputs corresponds with the measurements of the mobile device.

Data origin	Model 1		Model 2		Model 3	
	Inputs	Output	Inputs	Output	Inputs	Outputs
Fixed devices	Temperature	Indoor temperature	Humidity	Relative humidity	CO ₂ levels	CO ₂ concentration
Weather	Temperature		Humidity		—	
	Irradiance		Irradiance		—	
	Hour day		Hour day		Hour day	
Temporal	Day week	Year day	Day week	Year day	Day week	Year day
	Year day		Year day		Year day	
Distances to	Mobile device		Mobile device		Mobile device	

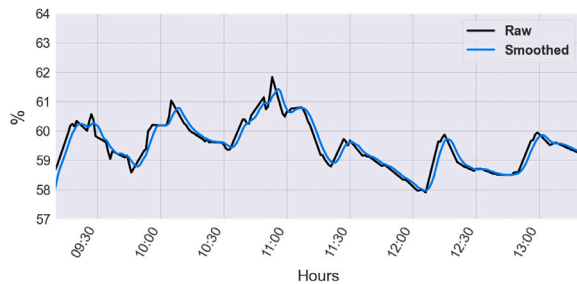


Fig. 5. Smoothing process applied to relative humidity data collected by the mobile device.

Table 4

Number of malfunctioning fixed devices supported by the built models based on the total fixed devices considered in the models.

	N° of fixed devices	N° of non-operating devices supported
Range 1	3–4	1
Range 2	5–6	2
Range 3	7–8	3

3. Results and discussion

A comparison between a set of original models, and their reduced versions obtained through the NSGA-III, using the XGBoost algorithm, is presented in this paper. The analysis focuses on measuring the changes in errors and computational time required for model training cause by the simplification. The available monitored data consists of almost 7 months of minute observations (taking into account that missing values represent approximately a 13% of the total). First, Section 3.1 presents the NSGA-III results by selecting the optimal

number of wall-mounted devices and decision trees in the XGBoost model. Second, Section 3.2 shows the average accuracy, together with the average computational time spent training, obtained by the built XGBoost models through a positional cross-validation. This process consists of performing interpolations at each of the positions through which the mobile device passed and collected data but previously extracting this data from the training sample. In addition, as a temporal analysis, interpolations produced over several days at a position where the original and reduced model can be studied together (missing values do not interfere with the reduced model estimations) are analysed together with the monitored values.

3.1. Model optimisation

An optimal selection of wall-mounted devices and the number of decision trees of the XGBoost models used in the reduced model are obtained with the NSGA-III. The CV(RMSE) given for each of the three analysed variables together with the number of fixed devices taking into account as model inputs are the four objective functions that the multi-objective genetic algorithm aims to minimise. Table 5 presents the parameters of NSGA-III used to perform the optimisation and the options space in which the genetic algorithm can operate. For simplicity, the calculation of the errors considered in the optimisation problem is based on the estimations produced by the models at three specific spatially and temporally separated positions (see Fig. 4).

The optimisation results are shown in Fig. 6 and represent the relation between the four objective functions considered in this study and the relative importance of each of the wall-mounted devices in this analysis. The results corresponds to a Pareto front in three dimensions (average CV(RMSE) values for the three analysed variables) separating by colour the points according to the number of wall-mounted devices considered (see left of Fig. 6). Additionally, the relative frequency

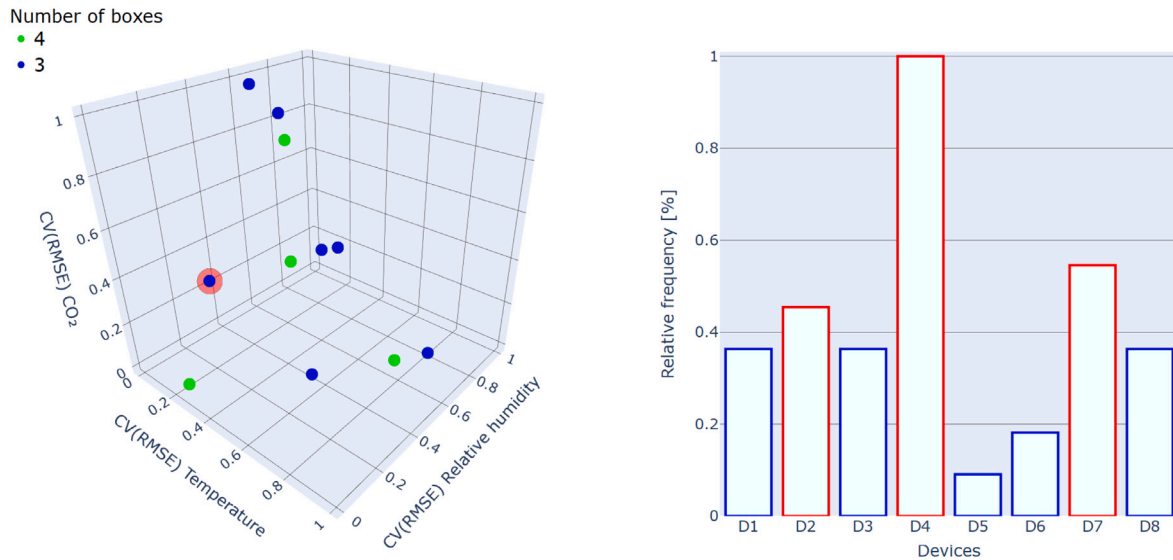


Fig. 6. NSGA-III results. On the left, the Pareto front (four objectives) obtained with the selected optimum point marked in red. The four objectives are represented: number of devices and the CV(RMSE), normalised, for each of the variables. On the right, the frequency in the optimal solution of each of the fixed devices and the final selection marked in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5

Parameters considered in the optimisation process. NSGA-III parameters are related with the performance of the genetic algorithm. Model options refer on the one hand to the XGBoost hyperparameters and, on the other hand, to the model complexity.

NSGA-III	Value	Model	Value
Population	100	N° trees	20 - 240 (20 by 20)
Mutation	0.3	N° fixed devices	3-8
Crossover	0.9		
N° generations	50		
Decision function	<i>pseudo-weights</i>		

of each of the fixed devices observed in the Pareto front points is presented on the right side of Fig. 6. The selection of the optimum is carried out using the *pseudo-weights* function considering homogeneous weights (0.25 for each of the objective functions). Thus, the optimum selected represents the use of three wall-mounted devices (D2, D4 and D7) and the following number of regression trees: 20 for temperature, 20 for relative humidity and 240 for CO₂ levels.

The frequency analysis points out that the devices D2, D4 and D7 are the most common ones among the solutions obtained with the NSGA-III (the same as those selected by the algorithm as the best combination). This means that these devices collect the most significant data to the analysis presented. In this case, the selected wall-mounted devices are located at the northeast, centre and southwest areas of the building (see Fig. 4). This is consistent as the selected devices are regularly distributed throughout the analysed space.

In addition, the reliability of the results are tested in Fig. 7 through the *centre distance* indicator (Eq. (5)). Considering the data shown in Table 2 the number of evaluations performed by the genetic algorithm to build the Pareto front is 5000 but the total number of possible options to test would be 289080 (very computationally costly). Furthermore, Fig. 7 shows that after generation 20 the values reached by the combinations tested by NSGA-III in relation to the analysed objectives stabilise.

3.2. Model performance

In this section, a comparison between the original models and their reduced versions, built according to the results of the optimisation conducted by the NSGA-III, is performed. This study compares the performance of each one based on a positional cross-validation (spatial

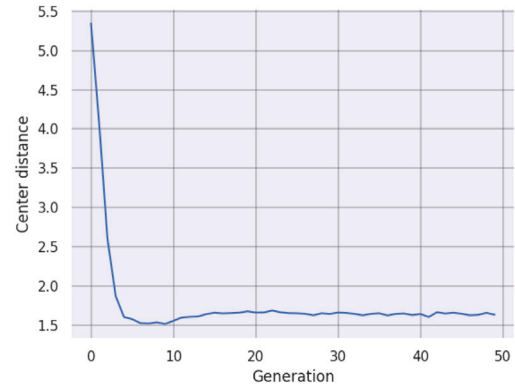


Fig. 7. NSGA-III convergence based on the centre distance indicator. The x-axis represents the different generations performed by the algorithm.

analysis) and their capacity to replicate the reality throughout several days (temporal analysis).

3.2.1. Spatial analysis

A positional cross-validation studies the average accuracy of the models by interpolating the variables of interest at each location of the mobile device (previously removing the data of the analysed position from the training sample). Thus, the efficiency of the models is analysed through several months considering different situations such as: occupancy patterns or weather conditions. The metrics considered were the NMBE (Eq. (6)), the CV(RMSE) (Eq. (7)) and the computational time, in seconds, required to be trained (see Table 6). In order to make the results comparable, all the numerical results are obtained using a server with a INTEL Core i7-8700K CPU, two NVIDIA Titan V GPUs and 64 GB of RAM.

The training process followed in this analysis was conducted in the following way:

1. The entire sample is divided into three sub-samples: training, test and validation. The test and validation samples each represent a specific position of the mobile device and the training sample represents the rest of the positions. The test sample is used to evaluate the training and select the best model hyperparameters

Table 6

Numerical results of the interpolations of the three studied variables in the analysed building. The mean NMBE, CV(RMSE) and computation time (in seconds), needed to train each of the models, obtained by positional cross-validation are presented with their standard deviations (SDs).

Model	Variables	Regression trees	NMBE [%]	SD	CV(RMSE) [%]	SD	c.t. [sec]
Original models	Temperature	200	-1.00	2.68	4.69	1.65	4.63
	Relative humidity	240	0.08	4.59	6.09	2.81	6.00
	CO ₂	60	-2.19	7.69	10.57	4.73	1.88
Reduced models	Temperature	20	0.17	3.48	5.52	1.71	0.99
	Relative humidity	20	0.97	4.41	5.73	3.23	1.13
	CO ₂	240	-2.20	7.37	9.82	4.51	1.02

and the validation sample is where the final performance of the model is evaluated.

2. To avoid overfitting problems, the training is performed taking into account an early stopping technique. If the performance of the model after certain iterations (in this case 20) does not improve or stagnates the training is stopped and the best iteration is considered (Nguyen-Sy et al., 2020).
3. During the training process the performance of the model is represented by the RMSE; being consistent with one of the metrics used in this study.
4. Additionally, the number of regression trees used in the original models are selected using a grid search approach based on 5-fold cross-validation (Muraier & Specht, 2018).
5. Repeat from step 1 but changing the positions selected to build the test and validation sample until all positions have been considered for the validation sample.

In the case of temperatures (model 1 in Table 3) and considering the original model, the average NMBE obtained in the entire sample is -1.00% with a standard deviation of ± 2.68 . In addition, the average CV(RMSE) is 4.69% showing a standard deviation of ± 1.65 (see Table 6). Taking into account its reduced version, the average NMBE becomes 0.17%, with a standard deviation of ± 3.48 , and the average CV(RMSE) becomes 5.52% showing a similar standard deviation of ± 1.71 . In the case of the average computational times needed by the models to train, it can be observed in Table 6 that the original model require more than 4 s and the reduced version less than 1 s (a 78% less). In the case of relative humidity (model 2 in Table 3), the original model shows an average NMBE of 0.08% with a standard deviation of ± 4.59 and an average CV(RMSE) of 6.09% with a variability of ± 2.81 . However, the reduced version obtained an average NMBE of 0.97% showing a standard deviation of ± 4.41 and an average CV(RMSE) of 5.73% with a variability of ± 3.23 (see Table 6). Regarding the average computational times, the reduced model demonstrates that, on average, needs a 81% less than the original model (1.13 and 6.00 s, respectively). Lastly, Table 6 also presents the average results of the models built to estimate CO₂ concentration (model 3 in Table 3). The average NMBE and the average CV(RMSE) yielded by the original model are -2.19% and 10.57%, respectively. Their standard deviations are ± 7.69 and ± 4.73 . In contrast, the reduced model shows an average NMBE of -2.20% with a variability of ± 7.37 , and an average CV(RMSE) of 9.82% with a standard deviation of ± 4.51 . Moreover, the reduced model shows an average computational time of 1.02 s while the original model require 1.88 s (a 46% less). In this case, the results are less accurate than for the other variables studied mainly due to the higher variability for short periods of time in the measurements. CO₂ concentration does not show such a strong inertia as in the case of temperature or relative humidity resulting in a less accuracy in the sensor used (see Table 2). The dependence on variables that are more difficult to monitor and have not been taken into account, such as occupancy rate or equipment use, also affects these results.

Consequently, in view of the above detailed results, the reduced models maintain low, or even reduce, the relative errors analysed, and although the variability of the results is slightly increased, the computational times are significantly reduced. Thus, the usefulness of NSGA-III for optimising the number of devices needed to efficiently estimate the IEQ of a building is demonstrated.

3.2.2. Temporal analysis

Once the overall performance of the models has been tested, a comparison between the temporal performance on specific days of the original model and the reduced model is presented (see Fig. 8). The time period studied is from 10 March to 13 March 2021 (only work days are considered in this analysis) and represents the location of the mobile device where the best aggregate results are obtained in the previous analysis.

It can be seen that for all variables, in general, both the original and the reduced model are able to replicate the actual behaviour pattern. Additionally, it can be observed that each of the studied variables shows a different variability in the measurements. Specifically, the CO₂ concentration values collected show the highest variability, resulting in less accurate results. As with the spatial analysis, the reduced version yielded similar results to the original model. In Fig. 8 it can be seen that both models present CV(RMSE) values around 4% for temperature and relative humidity and around 6.5% for CO₂ levels. Therefore, the higher training speed offered by the reduced models is not compromised by a lower accuracy of the estimates.

The optimised models show that it is not required the use of all wall-mounted devices to obtain accurate IEQ estimations in a large multi-zone open space. Both the errors and the variability in the obtained results are kept at similar values providing much lower computational times in the training process. Moreover, taking into account the reduced versions, the errors obtained by the built models are in the same range or even lower than those presented in other previous studies that also try to estimate the indoor environmental conditions inside buildings (Costantino et al., 2018; Pantazaras, Lee, Santamouris, & Yang, 2016; Rouchier, Rabouille, & Oberlé, 2018; Tagliabue, Re Cecconi, Rinaldi, & Ciribini, 2021).

4. Conclusions

A new application of multi-objective genetic algorithms and a new method for estimating the indoor environment conditions in a multi-zone open space within an existing building using the XGBoost algorithm are presented in this paper. The main conclusions of this research are:

- The optimised models built in this study allow the control of variables related with the indoor environment quality of a building every minute considering the possibility of malfunctioning monitoring devices.
- The results obtained show that, based on low-cost sensors measurements, the built models are able to accurately estimate the variables selected to summarise the IEQ of a building and that only three wall-mounted devices are required.
- The results demonstrated that, on the one hand, the built XGBoost models are able to efficiently estimate the IEQ of a building. The average CV(RMSE) they yield are below 5% for temperature, below 6% for relative humidity and below 10% for CO₂ levels. On the other hand, the reduced models are able to obtain similar results to the original ones in terms of relative errors by significantly reducing the installation complexity and the computational times required to train the model.

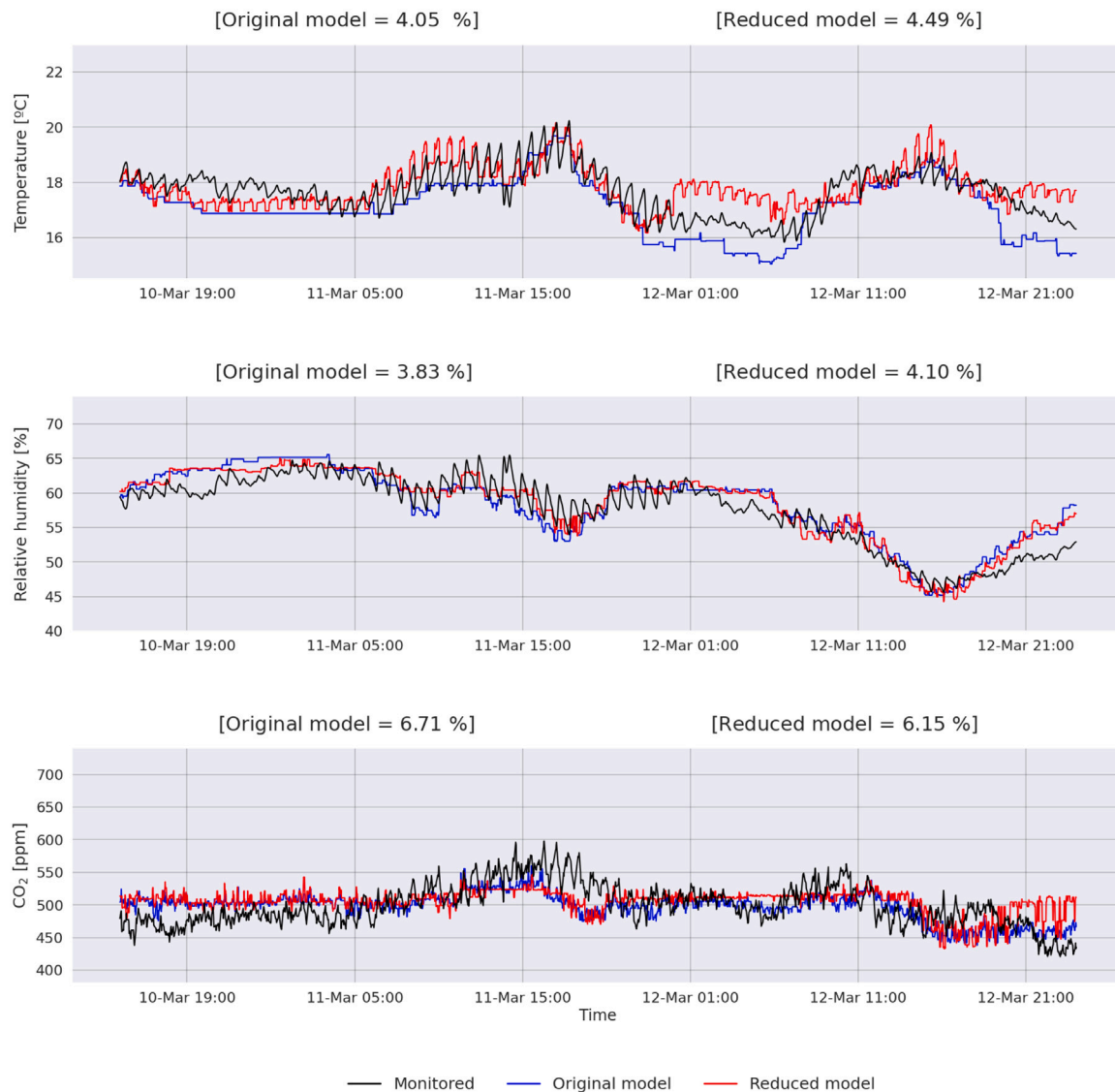


Fig. 8. Results of the interpolations carried out on specific days and in a unknown position by the models. In addition, the CV(RMSE) obtained by each of the models is presented on the graphs.

- The main contribution of this research is the application of mathematical models, whose internal structures are optimised through a multi-objective genetic algorithm, to accurately estimate the IEQ of a building in real time. The use of the NSGA-III, in addition to optimising the machine learning hyperparameters, has proven to be useful in reducing the number of fixed devices required to accurately interpolate data. Knowing precisely the optimal number of monitoring devices and the best locations to obtain accurate interpolations allows for less interference in the normal activities of building users. Moreover, the presented methodology can contribute to improve the monitoring of indoor environmental conditions and the detection of unexpected behaviours in the analysed variables. In this way, the estimations provided by the optimised model can be used to assess the building performance, detect flaws in the envelope or identify deficiencies in the cooling or heating system.
- The presented methodology has advantages over classic models as continuous monitoring due to its high cost or simulations that require a great expertise in the field and the control of many different parameters. In addition, previous studies based on machine learning models focus on small spaces, consider only fixed devices or require expensive equipment. In this case, the

system presented, produces equally accurate interpolations (better in some cases) with low-cost sensors, faster than traditional simulations and without the need for monitoring variables such as air exchange rate or occupancy rate (more complex and costly to monitor) and human intervention once the model has been trained.

- This method is applicable to other buildings since the case study is a large open space (in smaller rooms the problem would be simpler) and the variables selected to summarise the IEQ of a building are normally monitored or can even be substituted by others.
- A concrete future line of research is to extend the data collection to obtain a more extensive data sample. The length of the monitoring period is the main limitation of this study. Assuming that the available data corresponds to one year, all seasons and more positions along the analysed space would be considered to train the machine learning models. Thus, the model would have heterogeneous information to be more consistent. Other lines of future research could focus on the one hand, on the use of building design parameters as auxiliary inputs to the model or the study of more complex deep learning models that can replicate more accurately the evolution of the studied variables. On the

other hand, different and more accurate sensors could be taken into account in addition to other IEQ-related variables, such as radiant temperature, air exchange rate or illuminance, to improve IEQ analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements and funding

This research was partially supported by the Ministry of Science, Innovation and Universities of the Spanish government under the SMARTHERM project (RTI2018-096296-B-C2). The authors also want to thank the Ministry of Science, Innovation and Universities, Spain (grant FPU19/01187). Funding for open access charge: Universidade de Vigo/CISUG, Spain.

References

- Alavi, H., Forcada, N., Bortolini, R., & Edwards, D. J. (2021). Enhancing occupants' comfort through BIM-based probabilistic approach. *Automation in Construction*, 123, Article 103528. <http://dx.doi.org/10.1016/j.autcon.2020.103528>, URL <https://www.sciencedirect.com/science/article/pii/S0926580520311080>.
- Andargie, M. S., & Azar, E. (2019). An applied framework to evaluate the impact of indoor office environmental factors on occupants' comfort and working conditions. *Sustainable Cities and Society*, 46, Article 101447. <http://dx.doi.org/10.1016/j.scs.2019.101447>, URL <https://www.sciencedirect.com/science/article/pii/S2210670718307674>.
- American Society of Heating, Refrigerating and Air-Conditioning Engineers (2011). *ASHRAE guideline 10-2011, interactions affecting the achievement of acceptable indoor environments*. ASHRAE.
- Awadh, O. (2017). Sustainability and green building rating systems: LEED, BREEAM, GSAS and Estidama critical analysis. *Journal of Building Engineering*, 11, 25–29. <http://dx.doi.org/10.1016/j.jobe.2017.03.010>, URL <https://www.sciencedirect.com/science/article/pii/S2352710216301152>.
- Babajide Mustapha, I., & Saeed, F. (2016). Bioactive molecule prediction using extreme gradient boosting. *Molecules*, 21(8), <http://dx.doi.org/10.3390/molecules21080983>, URL <https://www.mdpi.com/1420-3049/21/8/983>.
- Babu, C. N., & Reddy, B. E. (2014). A moving-average filter based hybrid ARIMA-ANN model for forecasting time series data. *Applied Soft Computing*, 23, 27–38. <http://dx.doi.org/10.1016/j.asoc.2014.05.028>, URL <https://www.sciencedirect.com/science/article/pii/S1568494614002555>.
- Blank, J., & Deb, K. (2020). Pymoo: Multi-objective optimization in python. *IEEE Access*, 8, 89497–89509. <http://dx.doi.org/10.1109/access.2020.2990567>.
- Campos, G., Dugardin, F., Yalaoui, F., & Kelly, R. (2016). A NSGA-II and NSGA-III comparison for solving an open shop scheduling problem with resource constraints. *IFAC-PapersOnLine*, 49(12), 1272–1277. <http://dx.doi.org/10.1016/j.ifacol.2016.07.690>, 8th IFAC Conference on Manufacturing Modelling, Management and Control MIM 2016. URL <https://www.sciencedirect.com/science/article/pii/S2405896316309661>.
- Carmona, P., Climent, F., & Momparler, A. (2019). Predicting failure in the U.S. banking sector: An extreme gradient boosting approach. *International Review of Economics and Finance*, 61, 304–323. <http://dx.doi.org/10.1016/j.iref.2018.03.008>, URL <https://www.sciencedirect.com/science/article/pii/S1059056017306950>.
- Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *KDD '16, Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794). New York, NY, USA: Association for Computing Machinery, <http://dx.doi.org/10.1145/2939672.2939785>.
- Chen, Z., Li, C., & Sun, W. (2020). Bitcoin price prediction using machine learning: An approach to sample dimension engineering. *Journal of Computational and Applied Mathematics*, 365, Article 112395. <http://dx.doi.org/10.1016/j.cam.2019.112395>, URL <https://www.sciencedirect.com/science/article/pii/S037704271930398X>.
- Chen, J., Zhao, F., Sun, Y., & Yin, Y. (2020). Improved XGBoost model based on genetic algorithm. *International Journal of Computer Applications in Technology*, 62(3), 240–245.
- Costantino, A., Fabrizio, E., Ghigini, A., & Bariani, M. (2018). Climate control in broiler houses: A thermal model for the calculation of the energy use and indoor environmental conditions. *Energy and Buildings*, 169, 110–126. <http://dx.doi.org/10.1016/j.enbuild.2018.03.056>, URL <https://www.sciencedirect.com/science/article/pii/S0378778817338598>.
- Cui, Z., Chang, Y., Zhang, J., Cai, X., & Zhang, W. (2019). Improved NSGA-III with selection-and-elimination operator. *Swarm and Evolutionary Computation*, 49, 23–33. <http://dx.doi.org/10.1016/j.swevo.2019.05.011>, URL <https://www.sciencedirect.com/science/article/pii/S2210650218309787>.
- Elnaklah, R., Walker, I., & Natarajan, S. (2021). Moving to a green building: Indoor environment quality, thermal comfort and health. *Building and Environment*, 191, Article 107592. <http://dx.doi.org/10.1016/j.buildenv.2021.107592>, URL <https://www.sciencedirect.com/science/article/pii/S0360132321000081>.
- Franco, A., & Leccese, F. (2020). Measurement of CO2 concentration for occupancy estimation in educational buildings with energy efficiency purposes. *Journal of Building Engineering*, 32, Article 101714. <http://dx.doi.org/10.1016/j.jobe.2020.101714>.
- Frontczak, M., & Wargocki, P. (2011). Literature survey on how different factors influence human comfort in indoor environments. *Building and Environment*, 46(4), 922–937. <http://dx.doi.org/10.1016/j.buildenv.2010.10.021>, URL <https://www.sciencedirect.com/science/article/pii/S0360132310003136>.
- de Galicia, X. (2021). Meteogalicia. (accessed on 18 June 2021). URL <https://www.meteogalicia.gal/web/RSS/rssIndex.action>.
- Ganesh, G. A., Sinha, S. L., & Verma, T. N. (2020). Numerical simulation for optimization of the indoor environment of an occupied office building using double-panel and ventilation radiator. *Journal of Building Engineering*, 29, Article 101139. <http://dx.doi.org/10.1016/j.jobe.2019.101139>, URL <https://www.sciencedirect.com/science/article/pii/S2352710219306783>.
- Gao, X., Fan, S., Li, X., Guo, Z., Zhang, H., Peng, Y., et al. (2017). An improved XGBoost based on weighted column subsampling for object classification. In *2017 4th international conference on systems and informatics (ICSAI)* (pp. 1557–1562). <http://dx.doi.org/10.1109/ICSAI.2017.8248532>.
- Gilani, S., Montazeri, H., & Blocken, B. (2016). CFD simulation of stratified indoor environment in displacement ventilation: Validation and sensitivity analysis. *Building and Environment*, 95, 299–313. <http://dx.doi.org/10.1016/j.buildenv.2015.09.010>, URL <https://www.sciencedirect.com/science/article/pii/S0360132315301189>.
- Gong, M., Bai, Y., Qin, J., Wang, J., Yang, P., & Wang, S. (2020). Gradient boosting machine for predicting return temperature of district heating system: A case study for residential buildings in Tianjin. *Journal of Building Engineering*, 27, Article 100950. <http://dx.doi.org/10.1016/j.jobe.2019.100950>, URL <https://www.sciencedirect.com/science/article/pii/S2352710219307570>.
- Guerra-Santin, O., & Tweed, C. A. (2015). In-use monitoring of buildings: An overview of data collection methods. *Energy and Buildings*, 93, 189–207. <http://dx.doi.org/10.1016/j.enbuild.2015.02.042>, URL <https://www.sciencedirect.com/science/article/pii/S0378778815001474>.
- Han, Z., Gao, R. X., & Fan, Z. (2012). Occupancy and indoor environment quality sensing for smart buildings. In *2012 IEEE international instrumentation and measurement technology conference proceedings* (pp. 882–887). <http://dx.doi.org/10.1109/I2MTC.2012.6229557>.
- Hesaraki, A., Myhren, J. A., & Holmberg, S. (2015). Influence of different ventilation levels on indoor air quality and energy savings: A case study of a single-family house. *Sustainable Cities and Society*, 19, 165–172. <http://dx.doi.org/10.1016/j.scs.2015.08.004>, URL <https://www.sciencedirect.com/science/article/pii/S2210670715300251>.
- Huang, Y., Miles, H., & Zhang, P. (2020). A sequential modelling approach for indoor temperature prediction and heating control in smart buildings. *arXiv preprint arXiv:2009.09847*.
- Jiang, Q., Wang, L., Hei, X., Yu, G., Lin, Y., & Lu, X. (2016). MOEA/D-ARA+SBX: A new multi-objective evolutionary algorithm based on decomposition with artificial raindrop algorithm and simulated binary crossover. *Knowledge-Based Systems*, 107, 197–218. <http://dx.doi.org/10.1016/j.knsys.2016.06.007>, URL <https://www.sciencedirect.com/science/article/pii/S0950705116301708>.
- Jiao, Z., Sun, Y., Hong, Y., Parker, T., Hu, P., Mannan, M. S., et al. (2020). Development of flammable property quantitative property-consequence relationship (QPCR) models using extreme gradient boosting. *Industrial and Engineering Chemistry Research*, XXXX, <http://dx.doi.org/10.1021/acs.iecr.0c02822>.
- Jin, M., Liu, S., Tian, Y., Lu, M., Schiavon, S., & Spanos, C. (2017). Indoor environmental quality monitoring by autonomous mobile sensing. In *BuildSys '17, Proceedings of the 4th ACM international conference on systems for energy-efficient built environments*. New York, NY, USA: Association for Computing Machinery, <http://dx.doi.org/10.1145/3137133.3137158>.
- Komasi, M., & Goudarzi, H. (2020). Multi-objective optimization of groundwater monitoring network using a probability Pareto genetic algorithm and entropy method (case study: Silakhor plain). *Journal of Hydroinformatics*, 23(1), 136–150. <http://dx.doi.org/10.2166/hydro.2020.061>, [arXiv:https://iwaponline.com/jh/article-pdf/23/1/136/846238/jh0230136.pdf](https://iwaponline.com/jh/article-pdf/23/1/136/846238/jh0230136.pdf).
- Kumari, P., & Toshniwal, D. (2021). Extreme gradient boosting and deep neural network based ensemble learning approach to forecast hourly solar irradiance. *Journal of Cleaner Production*, 279, Article 123285. <http://dx.doi.org/10.1016/j.jclepro.2020.123285>, URL <https://www.sciencedirect.com/science/article/pii/S0959652620333308>.
- Li, H., Deb, K., Zhang, Q., Suganthan, P., & Chen, L. (2019). Comparison between MOEA/D and NSGA-III on a set of novel many and multi-objective benchmark problems with challenging difficulties. *Swarm and Evolutionary Computation*, 46, 104–117. <http://dx.doi.org/10.1016/j.swevo.2019.02.003>, URL <https://www.sciencedirect.com/science/article/pii/S2210650218307016>.

- Li, Y.-H., Sheng, Z., Zhi, P., & Li, D. (2019). Multi-objective optimization design of anti-rolling torsion bar based on modified NSGA-III algorithm. *International Journal of Structural Integrity, ahead-of-print*, <http://dx.doi.org/10.1108/IJSI-03-2019-0018>.
- Lu, H., Cheng, F., Ma, X., & Hu, G. (2020). Short-term prediction of building energy consumption employing an improved extreme gradient boosting model: A case study of an intake tower. *Energy*, 203, Article 117756. <http://dx.doi.org/10.1016/j.energy.2020.117756>, URL <https://www.sciencedirect.com/science/article/pii/S036054422030863X>.
- Ma, N., Aviv, D., Guo, H., & Brahm, W. W. (2021). Measuring the right factors: A review of variables and models for thermal comfort and indoor air quality. *Renewable and Sustainable Energy Reviews*, 135, Article 110436. <http://dx.doi.org/10.1016/j.rser.2020.110436>, URL <https://www.sciencedirect.com/science/article/pii/S1364032120307231>.
- Martínez, S., Eguía, P., Granada, E., Moazami, A., & Hamdy, M. (2020). A performance comparison of multi-objective optimization-based approaches for calibrating white-box building energy models. *Energy and Buildings*, 216, Article 109942. <http://dx.doi.org/10.1016/j.enbuild.2020.109942>, URL <https://www.sciencedirect.com/science/article/pii/S0378778819336850>.
- Martínez-Comesaña, M., Febrero-Garrido, L., Granada-Álvarez, E., Martínez-Torres, J., & Martínez-Mariño, S. (2020). Heat loss coefficient estimation applied to existing buildings through machine learning models. *Applied Sciences*, 10(24), <http://dx.doi.org/10.3390/app10248968>, URL <https://www.mdpi.com/2076-3417/10/24/8968>.
- Martínez-Comesaña, M., Febrero-Garrido, L., Troncoso-Pastoriza, F., & Martínez-Torres, J. (2020). Prediction of building's thermal performance using LSTM and MLP neural networks. *Applied Sciences*, 10(21), <http://dx.doi.org/10.3390/app10217439>, URL <https://www.mdpi.com/2076-3417/10/21/7439>.
- Martínez-Comesaña, M., Ogando-Martínez, A., Troncoso-Pastoriza, F., López-Gómez, J., Febrero-Garrido, L., & Granada-Álvarez, E. (2021). Use of optimised MLP neural networks for spatiotemporal estimation of indoor environmental conditions of existing buildings. *Building and Environment*, 205, Article 108243. <http://dx.doi.org/10.1016/j.buildenv.2021.108243>, URL <https://www.sciencedirect.com/science/article/pii/S0360132321006429>.
- Martínez-Mariño, S., Eguía-Oller, P., Granada-Álvarez, E., & Erkoreka-González, A. (2021). Simulation and validation of indoor temperatures and relative humidity in multi-zone buildings under occupancy conditions using multi-objective calibration. *Building and Environment*, 200, Article 107973. <http://dx.doi.org/10.1016/j.buildenv.2021.107973>, URL <https://www.sciencedirect.com/science/article/pii/S0360132321003772>.
- Murauer, B., & Specht, G. (2018). Detecting music genre using extreme gradient boosting. In *WWW '18, Companion proceedings of the the web conference 2018* (pp. 1923–1927). Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, <http://dx.doi.org/10.1145/3184558.3191822>.
- Nguyen-Sy, T., Wakim, J., To, Q.-D., Vu, M.-N., Nguyen, T.-D., & Nguyen, T.-T. (2020). Predicting the compressive strength of concrete from its compositions and age using the extreme gradient boosting method. *Construction and Building Materials*, 260, Article 119757. <http://dx.doi.org/10.1016/j.conbuildmat.2020.119757>, URL <https://www.sciencedirect.com/science/article/pii/S0950061820317621>.
- NOAA (2021a). The global forecast system (GFS) - global spectral model (GSM). (accessed on 18 June 2021). URL https://www.emc.ncep.noaa.gov/emc/pages/numerical_forecast_systems/gfs/documentation.php.
- NOAA (2021b). NOAA operational model archive and distribution system (NOMADS). (accessed on 18 June 2021). URL <https://nomads.ncep.noaa.gov/>.
- Pantazaras, A., Lee, S. E., Santamouris, M., & Yang, J. (2016). Predicting the CO₂ levels in buildings using deterministic and identified models. *Energy and Buildings*, 127, 774–785. <http://dx.doi.org/10.1016/j.enbuild.2016.06.029>, URL <https://www.sciencedirect.com/science/article/pii/S0378778816305187>.
- Park, J., Loftness, V., Aziz, A., & Wang, T.-H. (2019). Critical factors and thresholds for user satisfaction on air quality in office environments. *Building and Environment*, 164, Article 106310. <http://dx.doi.org/10.1016/j.buildenv.2019.106310>, URL <https://www.sciencedirect.com/science/article/pii/S0360132319305207>.
- Parvizi, E., & Rezvani, M. H. (2020). Utilization-aware energy-efficient virtual machine placement in cloud networks using NSGA-III meta-heuristic approach. *Cluster Computing*, 23(4), 2945–2967. <http://dx.doi.org/10.1007/s10586-020-03060-y>.
- Ritchie, H., & Roser, M. (2013). *Indoor air pollution*. Our World in Data, <https://ourworldindata.org/indoor-air-pollution>.
- Roberts, C., Edwards, D., Hosseini, M. R., Mateo-García, M., & Owusu-Manu, D.-G. (2019). Post-occupancy evaluation: a review of literature. *Engineering, Construction and Architectural Management*, 26, <http://dx.doi.org/10.1108/ECAM-09-2018-0390>.
- Rouchier, S., Rabouille, M., & Oberlé, P. (2018). Calibration of simplified building energy models for parameter estimation and forecasting: Stochastic versus deterministic modelling. *Building and Environment*, 134, 181–190. <http://dx.doi.org/10.1016/j.buildenv.2018.02.043>, URL <https://www.sciencedirect.com/science/article/pii/S036013231830115X>.
- Ruiz, G. R., & Bandera, C. F. (2017). Validation of calibrated energy models: Common errors. *Energies*, 10(10), <http://dx.doi.org/10.3390/en10101587>, URL <https://www.mdpi.com/1996-1073/10/10/1587>.
- Saini, J., Dutta, M., & Marques, G. (2020). Indoor air quality monitoring with IoT: Predicting PM10 for enhanced decision support. In *2020 international conference on decision aid sciences and application (DASA)* (pp. 504–508). <http://dx.doi.org/10.1109/DASA51403.2020.9317054>.
- Sang, B. (2021). Application of genetic algorithm and BP neural network in supply chain finance under information sharing. *Journal of Computational and Applied Mathematics*, 384, Article 113170. <http://dx.doi.org/10.1016/j.cam.2020.113170>, URL <https://www.sciencedirect.com/science/article/pii/S0377042720304611>.
- Sant'Anna, D., Dos Santos, P., Vianna, N., & Romero, M. (2018). Indoor environmental quality perception and users' satisfaction of conventional and green buildings in Brazil. *Sustainable Cities and Society*, 43, 95–110. <http://dx.doi.org/10.1016/j.scs.2018.08.027>, URL <https://www.sciencedirect.com/science/article/pii/S2210670718308606>.
- Sha, H., Xu, P., Hu, C., Li, Z., Chen, Y., & Chen, Z. (2019). A simplified HVAC energy prediction method based on degree-day. *Sustainable Cities and Society*, 51, Article 101698. <http://dx.doi.org/10.1016/j.scs.2019.101698>, URL <https://www.sciencedirect.com/science/article/pii/S2210670719304706>.
- Shan, X., & Yang, E.-H. (2020). Supervised machine learning of thermal comfort under different indoor temperatures using EEG measurements. *Energy and Buildings*, 225, Article 110305. <http://dx.doi.org/10.1016/j.enbuild.2020.110305>, URL <https://www.sciencedirect.com/science/article/pii/S0378778820311993>.
- Sharma, P. K., Mondal, A., Jaiswal, S., Saha, M., Nandi, S., De, T., et al. (2021). IndoAirSense: A framework for indoor air quality estimation and forecasting. *Atmospheric Pollution Research*, 12(1), 10–22.
- Shen, H., Hou, W., Zhu, Y., Zheng, S., Ainiwaer, S., Shen, G., et al. (2021). Temporal and spatial variation of PM_{2.5} in indoor air monitored by low-cost sensors. *Science of the Total Environment*, 770, Article 145304. <http://dx.doi.org/10.1016/j.scitotenv.2021.145304>, URL <https://www.sciencedirect.com/science/article/pii/S0048969721003703>.
- Shrestha, M., Rijal, H., Kayo, G., & Shukuya, M. (2022). An investigation on CO₂ concentration based on field survey and simulation in naturally ventilated Nepalese school buildings during summer. *Building and Environment*, 207, Article 108405. <http://dx.doi.org/10.1016/j.buildenv.2021.108405>, URL <https://www.sciencedirect.com/science/article/pii/S0360132321008027>.
- Steinemann, A., Wargocki, P., & Rismanchi, B. (2017). Ten questions concerning green buildings and indoor air quality. *Building and Environment*, 112, 351–358. <http://dx.doi.org/10.1016/j.buildenv.2016.11.010>, URL <https://www.sciencedirect.com/science/article/pii/S0360132316304346>.
- Tagliabue, L. C., Re Cecconi, F., Rinaldi, S., & Ciribini, A. L. C. (2021). Data driven indoor air quality prediction in educational facilities based on IoT network. *Energy and Buildings*, 236, Article 110782. <http://dx.doi.org/10.1016/j.enbuild.2021.110782>, URL <https://www.sciencedirect.com/science/article/pii/S0378778821000669>.
- Thongsuwan, S., Jaiyen, S., Padcharoen, A., & Agarwal, P. (2021). ConvXGB: A new deep learning model for classification problems based on CNN and XG-Boost. *Nuclear Engineering and Technology*, 53(2), 522–531. <http://dx.doi.org/10.1016/j.net.2020.04.008>, URL <https://www.sciencedirect.com/science/article/pii/S173857319308587>.
- Tien, P., Wei, S., Calautit, J. K., Darkwa, J., & Wood, C. (2020). A vision-based deep learning approach for the detection and prediction of occupancy heat emissions for demand-driven control solutions. *Energy and Buildings*, 226, Article 110386. <http://dx.doi.org/10.1016/j.enbuild.2020.110386>.
- Vileckova, S., Meciarova, L., Burdova, E. K., Katunská, J., Kosicanova, D., & Doroudiani, S. (2017). Indoor environmental quality of classrooms and occupants' comfort in a special education school in Slovak Republic. *Building and Environment*, 120, 29–40. <http://dx.doi.org/10.1016/j.buildenv.2017.05.001>, URL <https://www.sciencedirect.com/science/article/pii/S0360132317301749>.
- Wang, Y., Chen, C., Tao, Y., Wen, Z., Chen, B., & Zhang, H. (2019). A many-objective optimization of industrial environmental management using NSGA-III: A case of China's iron and steel industry. *Applied Energy*, 242, 46–56. <http://dx.doi.org/10.1016/j.apenergy.2019.03.048>, URL <https://www.sciencedirect.com/science/article/pii/S03606261919304568>.
- Wang, R., Lu, S., & Li, Q. (2019). Multi-criteria comprehensive study on predictive algorithm of hourly heating energy consumption for residential buildings. *Sustainable Cities and Society*, 49, Article 101623. <http://dx.doi.org/10.1016/j.scs.2019.101623>, URL <https://www.sciencedirect.com/science/article/pii/S2210670719309527>.
- Wang, S., Yan, C., & Xiao, F. (2012). Quantitative energy performance assessment methods for existing buildings. *Energy and Buildings*, 55, 873–888. <http://dx.doi.org/10.1016/j.enbuild.2012.08.037>, Cool Roofs, Cool Pavements, Cool Cities, and Cool World. URL <https://www.sciencedirect.com/science/article/pii/S0378778812004410>.
- Wu, J., Hou, Z., Shen, J., & Lian, Z. (2021). A method for the determination of optimal indoor environmental parameters range considering work performance. *Journal of Building Engineering*, 35, Article 101976. <http://dx.doi.org/10.1016/j.job.2020.101976>, URL <https://www.sciencedirect.com/science/article/pii/S2352710220336081>.
- Yang, Y., Liu, J., Wang, W., Cao, Y., & Li, H. (2021). Incorporating SLAM and mobile sensing for indoor CO₂ monitoring and source position estimation. *Journal of Cleaner Production*, 291, Article 125780. <http://dx.doi.org/10.1016/j.jclepro.2020.125780>, URL <https://www.sciencedirect.com/science/article/pii/S0959652620358261>.
- Yi, J.-H., Xing, L.-N., Wang, G.-G., Dong, J., Vasilakos, A., Alavi, A., et al. (2018). Behavior of crossover operators in NSGA-III for large-scale optimization problems. *Information Sciences*, 509, <http://dx.doi.org/10.1016/j.ins.2018.10.005>.

- Yin, X., Zhao, X., Lin, J., & Karcianas, A. (2020). Reliability aware multi-objective predictive control for wind farm based on machine learning and heuristic optimizations. *Energy*, 202, Article 117739. <http://dx.doi.org/10.1016/j.energy.2020.117739>, URL <https://www.sciencedirect.com/science/article/pii/S036054422030846X>.
- Yu, Z., Song, Y., Song, D., & Liu, Y. (2021). Spatial interpolation-based analysis method targeting visualization of the indoor thermal environment. *Building and Environment*, 188, Article 107484. <http://dx.doi.org/10.1016/j.buildenv.2020.107484>, URL <https://www.sciencedirect.com/science/article/pii/S0360132320308519>.
- Yu, X., Wang, Y., Wu, L., Chen, G., Wang, L., & Qin, H. (2020). Comparison of support vector regression and extreme gradient boosting for decomposition-based data-driven 10-day streamflow forecasting. *Journal of Hydrology*, 582, Article 124293. <http://dx.doi.org/10.1016/j.jhydrol.2019.124293>, URL <https://www.sciencedirect.com/science/article/pii/S0022169419310285>.
- Zhang, W., Wu, C., Zhong, H., Li, Y., & Wang, L. (2021). Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. *Geoscience Frontiers*, 12(1), 469–477. <http://dx.doi.org/10.1016/j.gsf.2020.03.007>, URL <https://www.sciencedirect.com/science/article/pii/S1674987120300669>.
- Zhu, Y., Liang, J., Chen, J., & Ming, Z. (2017). An improved NSGA-III algorithm for feature selection used in intrusion detection. *Knowledge-Based Systems*, 116, 74–85. <http://dx.doi.org/10.1016/j.knosys.2016.10.030>, URL <https://www.sciencedirect.com/science/article/pii/S0950705116304245>.
- Zuhaib, S., Manton, R., Griffin, C., Hajdukiewicz, M., Keane, M. M., & Goggins, J. (2018). An Indoor Environmental Quality (IEQ) assessment of a partially-retrofitted university building. *Building and Environment*, 139, 69–85. <http://dx.doi.org/10.1016/j.buildenv.2018.05.001>, URL <https://www.sciencedirect.com/science/article/pii/S0360132318302610>.