Surrogate Model Development for Naturally Ventilated Office Buildings

Marcelo Salles Olinger¹, Ana Paula Melo¹, Leticia Oliveira Neves², Roberto Lamberts¹
¹Federal University of Santa Catarina, Florianópolis, Brazil
²University of Campinas, Campinas, Brazil

Abstract
Building energy simulation tools are very helpful to achieve thermal performance for buildings. However, modeling can require much detail, specially related to input data. The use of machine learning to develop surrogate models can support architects and builders to get useful information on buildings thermal performance, in a fast and simple way. The aim of this study is to present a machine learning methodology to develop a surrogate model for naturally ventilated office buildings, using artificial neural networks. The output of the surrogate model is the Exceedance Hour Fraction (EHF), a thermal comfort indicator. The final surrogate model has 12 input parameters that can estimate thermal comfort for offices with a wide range of characteristics. The mean absolute error measured for the surrogate model was 0.04.

Introduction
Energy demand for cooling in buildings has increased more than three times from 1990 to 2016. The energy use for cooling in 2016 was 2 020 TWh, corresponding to 1/5 of the total energy use in the sector. The current baseline scenario for the year of 2050 indicates that the energy use for cooling will continue to grow, representing 37% of the increase in electricity use in buildings (IEA, 2018). This report also indicates that in countries with hot climate and emerging economies this issue is even more relevant. Only 8% of the 2.8 billions of people living in these countries today use artificial cooling. By 2050, the share of cooling in electricity system peak loads could get to 30,8% in Brazil. In India, this could get up to 44,1%.

An alternative solution for artificial cooling systems is the use of passive cooling strategies, such as natural ventilation (NV). In many countries, people occupy naturally ventilated buildings and expect for adaptive comfort standards (de Dear et al., 1997). This context indicates the importance of NV strategies for the mitigation of artificial cooling in buildings.

The potential of NV in office buildings has been studied in numerous cases, considering a variety of ventilation modes, such as night ventilation or mixed-mode (artificial cooling along with NV) (Yun et al., 2008; Yao et al., 2009; Carrilho da Graça and Linden, 2016; Elharidi et al., 2018; Pesic et al., 2018). Although the use of cross ventilation and shading devices are the predominant proposed solutions, certain characteristics of a building can result in different thermal performances, according to different combination of parameters and the climate characteristics. Therefore, analysis of NV in early design phases is crucial to achieve better thermal performance in buildings.

In early phases, the potential for optimization is significant, and any estimate of occupants comfort and thermal performance of the building has influence on decision making (Belleri et al., 2014; Roetzel et al., 2014). However, there is a wide number of possible solutions for thermal performance in buildings and to analyze all the possibilities in early design phases can be complex and time-consuming.

Building Energy Simulation (BES) can provide information on thermal performance of buildings using NV, but the development of such models require expertise and detailed input data, which may not be available in early design phases (Østergård et al., 2016). The analysis of different solutions for thermal performance in a simple and fast way is fundamental for early design phases. The use of a surrogate model could achieve similar results to the original simulation model, and allow architects to perform analysis and visualize the impact of design decisions (Alsaadani and Souza, 2019).

Machine learning has been widely used to develop surrogate models aiming to estimate energy efficiency in buildings. Melo et al. (2016) compared the multilinear regression method to artificial neural networks (ANN) to improve the accuracy of Brazilian surrogate models for labeling purposes. The results showed a significant improvement by applying ANN, instead of multiple linear regression. Rackes et al. (2016) developed a support vector machine model to estimate thermal comfort levels in commercial low-rise buildings in Brazil. The surrogate model shows the influence that changing the set of parameters can have on thermal comfort. Østergård et al. (2018) described an approach to explore multidimensional space in early
design stage through global sensitivity analysis. The work emphasizes how different parameter configurations should consider not just one, but a range of objectives, related to energy consumption, daylight, thermal comfort, cost and others. Wei (2013) presented a review on sensitivity analysis by comparing different methods and discussing the implications on the adoption of methods that do not consider colinearities and non-monotonic behavior. The global sensitivity analysis method proposed by Sobol’ (1993) identifies the input factors that can be fixed at any given value over their range of uncertainty without significantly reducing the output variance.

The aim of this paper is to present a machine learning methodology for naturally ventilated office buildings, to achieve a robust and accurate surrogate model that estimates thermal comfort. A database is used to identify the most common characteristics of naturally ventilated office buildings. Sensitivity Analysis (SA) is applied to understand the most influential parameters that should be taken into account on the surrogate model. ANN is adopted as the machine learning model.

Method
This section explains each step of the methodology proposed in this study. Figure 1 presents a flowchart with these steps.

![Figure 1: Steps taken for the development of the surrogate model.](image)

Database
Pereira and Neves (2018) developed a database with information about office buildings that work with mixed-mode ventilation in the city of Sao Paulo, Brazil. A total of 153 buildings constructed after 1995 were selected. A survey in loco was conducted in 50 of those buildings, to gather information about the geometry of the offices, characteristics of the windows and shading elements. The survey allowed to identify the most recurrent design characteristics of the buildings and their ranges. The buildings’ design information include: geometry (area, dimensions, number of floors); solar absorptance of roof and exterior walls and window-to-wall ratio (WWR). Figure 2 presents an example of how these data was explored.

![Figure 2: Histograms of WWR and area of the offices available in the database.](image)

Base model
The database provided information to build a large number of BES models, in order to explore its whole range of possible characteristics. It is also possible to widen the range of some features, specially on those more relevant to buildings thermal performance.

The dataset used to apply a sensitivity analysis and to develop the ANN was generated with simulations from the software EnergyPlus (2018), version 8.9. The office room was modeled as a rectangular single zone (Figure 3), as it allows fast simulations with different geometries and exposure condition of the surfaces. Although the model was developed as a single zone, the idea is to represent an office within the whole building, and this aspect had to be considered to define the boundary conditions of building surfaces that represented a wall, floor or ceiling adjacent to a building. To validate the single zone approach, a comparison between thermal zones considering the whole building and single zone models was conducted, based on a sample of 100 cases generated by Latin Hypercube Sampling (LHS).

The weather file of the city of Sao Paulo was obtained from INMET (2016).

The boundary conditions of walls and ceiling were modeled as either outdoors (in case it represented an external wall or a roof), or adiabatic (in case it represented a surface adjacent to other surfaces of the building). The floor’s boundary condition was mod-
eled as ground or adiabatic surface, if it represented superior floors.

Figure 3: 3D model example

EnergyPlus uses the algorithms from the Airflow Network (AFN) models (Walton, 1989), so the door of the office had to be accounted as a linkage on the network (Figure 4). Therefore, the door could not be modeled as adiabatic, even though it was adjacent to the corridor. This issue was overcome by considering the wall with the door as an outdoor surface element, with no wind and no sun exposure.

The wind pressure coefficient \( C_p \) is fundamental for the AFN simulation, since it defines the pressure differences between the building surfaces. \( C_p \) was based on the ASHRAE Handbook (2001), using the algorithm executed by EnergyPlus for high-rise buildings. Furthermore, \( C_p \) was manually input instead of automatically calculated by EnergyPlus, otherwise the software would estimate the pressure node outside the door as if it was outside the building. This issue was overcome by calculating an equivalent \( C_p \) (Equation 1) for the office door node, for each wind angle. The calculation was made by considering a building floor plan of six offices. Each door linkage was considered as a crack flow, and each window was considered as a large opening.

\[
C_{eq,\alpha} = \frac{P_{corr}}{P_{out}} \tag{1}
\]

Where \( C_{eq,\alpha} \) is the equivalent \( C_p \) for the angle \( \alpha \); \( P_{corr} \) is the pressure in the corridor; and \( P_{out} \) is the outdoor pressure. \( P_{out} \) is defined by EnergyPlus, based on the wind speed.

\( P_{corr} \) was calculated based on the AFN pressure balance, so it considers the entire airflow that comes to/from the corridor pressure point, and sum up the airflow to equal zero. The airflow across linkages considered as doors is calculated as a crack airflow, while the airflow across linkages considered as windows is calculated as a large opening. The pressure of each zone, related to the values of \( C_p \), was calculated with an equation solver algorithm, according to Equation 2, where \( N_d \) is the number of doors that link to the zone; \( N_w \) is the number of windows that link to the zone; \( C_{door} \) is the coefficient related to the discharge on the airflow through the doors; \( C_{window} \) is the coefficient related to the discharge on the airflow through the windows; \( P_{link,i} \) is the pressure in the zone linked by door \( i \); and \( C_{p,j} \) is the \( C_p \) on the surface of window \( j \).

\[
P_{zn} = \sum_{i=1}^{N_d} C_{door} \times (P_{zn} - P_{link,i}) + \sum_{j=1}^{N_w} C_{window} \times (P_{zn} - P_{out} \times C_{p,j}) \tag{2}
\]

Figure 4 presents an example of the relation between \( C_p \)'s, the pressures in the zone and outdoors, and the discharge related to the airflow linkages. For the example, subscripts "N" and "W" on \( C_p \) refer to north and west, respectively.

Figure 4: Relation between \( C_p \) and zone’s pressure.

Even though the model is a single zone, two characteristics that correspond to the building as a whole were considered: floor height and building ratio. Floor height is the distance of the zone’s floor from the ground. It is relevant to the wind speed consideration (calculated by EnergyPlus), thus to the NV performance. Building ratio is the ratio between width and length of the building, which is rectangular. This ratio is taken into account on the calculation of the \( C_p \). The algorithm that calculates \( C_p \) assumes that the building is rectangular, so that is a limitation of the proposed methodology.

The openings were modeled as a fenestration surface detailed object, with airflow determined by the AFN object Detailed Opening Component. Window size, opening factor and solar heat gain coefficient (SHGC) were varied according to the samples generated for the study. Window glazing was modeled using the Simple Glazing Object.

The shading device was considered as a horizontal, rectangular surface that hangs from the top of the window opening, as shown on Figure 3. The object used was Shading Building Detailed.

Occupation was defined from 8:00 am to 6:00 pm.
During occupation, the model considered thermal gains from people, electric equipment and lights. Unoccupied periods had no internal gains. Electric Equipment heat load was considered 150 W/person. Lights were defined as 10.50 W/m². The metabolic rate for occupation was defined as 120 W/person.

Thermal properties of the walls were defined by considering the walls exposed to outdoors as a two layers construction component. The inside layer in concrete, used to define the thermal capacity of the wall, and the outside layer in Expanded polystyrene (EPS), modeled using the object Material:NoMass. EPS was used to determine the wall’s transmittance.

Ventilation control was available only during occupation, and it was based on the Temperature mode, defined as 20 °C.

Table 1 presents the parameters analyzed, with minimum and maximum values. Some parameters were not numerical, these are: contact to ground (yes/no); exterior roof (yes/no); exterior walls exposure (1 wall and 1 window; 2 walls and 1 window; 2 walls and 2 windows), as shown on Figure 5. The exposure condition of walls and windows defines, besides the heat exchange with the outdoor, if whether or not there will be cross ventilation.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Minimum Value</th>
<th>Maximum Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area [m²]</td>
<td>20</td>
<td>100</td>
</tr>
<tr>
<td>Zone Ratio [-]</td>
<td>0.5</td>
<td>2.0</td>
</tr>
<tr>
<td>Zone Height [m]</td>
<td>2.4</td>
<td>3.2</td>
</tr>
<tr>
<td>Absorptance [-]</td>
<td>0.3</td>
<td>0.9</td>
</tr>
<tr>
<td>Shading [m]</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Azimuth [°]</td>
<td>0.0</td>
<td>359.9</td>
</tr>
<tr>
<td>Wall Transmittance [W/m²K]</td>
<td>0.5</td>
<td>4.4</td>
</tr>
<tr>
<td>Wall Thermal Capacity [kJ/m²K]</td>
<td>20</td>
<td>400</td>
</tr>
<tr>
<td>WWR [-]</td>
<td>0.1</td>
<td>0.6</td>
</tr>
<tr>
<td>Opening Factor [-]</td>
<td>0.1</td>
<td>1.0</td>
</tr>
<tr>
<td>Occupant Density</td>
<td>0.05</td>
<td>0.50</td>
</tr>
<tr>
<td>SHGC [-]</td>
<td>0.20</td>
<td>0.87</td>
</tr>
<tr>
<td>Floor Height [m]</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>Building ratio [-]</td>
<td>0.25</td>
<td>4.00</td>
</tr>
</tbody>
</table>

Table 1: Parameters with minimum and maximum values.

The range of values for those parameters were varied in order to explore the limits of possibilities. Even if some of the values were not found in the existing database of real buildings, parameters were explored as much as possible, aiming to achieve thermal comfort results considered acceptable for the majority of the occupants.

Thermal comfort

Thermal comfort in naturally ventilated offices is determined based on the adaptive comfort method defined in ASHRAE Standard 55 (2017). The adaptive method determines the upper and lower limits for the operative temperature of a zone, that guarantees a certain range of acceptability among the occupants. These temperature limits are based on the mean outdoor temperature. Table 2 presents the monthly mean temperatures for the city of Sao Paulo (INMET, 2016). $T_{upper}$ is the operative temperature upper limit for 80% of acceptability.

Table 2: Outdoor mean temperatures of Sao Paulo.

<table>
<thead>
<tr>
<th>Month</th>
<th>Mean Temperature [°C]</th>
<th>$T_{upper}$ [°C]</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>21.16</td>
<td>28.66</td>
</tr>
<tr>
<td>February</td>
<td>22.35</td>
<td>29.03</td>
</tr>
<tr>
<td>March</td>
<td>21.67</td>
<td>28.82</td>
</tr>
<tr>
<td>April</td>
<td>20.76</td>
<td>28.54</td>
</tr>
<tr>
<td>May</td>
<td>17.45</td>
<td>27.51</td>
</tr>
<tr>
<td>June</td>
<td>16.77</td>
<td>27.30</td>
</tr>
<tr>
<td>July</td>
<td>17.34</td>
<td>27.48</td>
</tr>
<tr>
<td>August</td>
<td>18.28</td>
<td>27.77</td>
</tr>
<tr>
<td>September</td>
<td>17.68</td>
<td>27.58</td>
</tr>
<tr>
<td>October</td>
<td>20.51</td>
<td>28.46</td>
</tr>
<tr>
<td>November</td>
<td>20.15</td>
<td>28.35</td>
</tr>
<tr>
<td>December</td>
<td>20.87</td>
<td>28.37</td>
</tr>
</tbody>
</table>

The adopted thermal comfort metric was the Exceedance Hour Fraction (EHF). It is the fraction of occupied hours within the thermal zone, where the operative temperature is above the upper limit for 80% of acceptability, according to ASHRAE Standard 55 (2017). For each hour of the year with occupancy in the zone, the operative temperature is compared to the upper limit temperature and a value of 0 or 1 is given to the indicator $E_{hot}$, according to Equation 3.

$$E_{hot,i} = \begin{cases} 1, & T_{op,i} > T_{upper,i} \\ 0, & T_{op,i} \leq T_{upper,i} \end{cases}$$

Figure 5: Exposition of walls. The mirrored image is also considered.
Where \( E_{\text{hot},i} \) is the binary indicator of operative temperature above the upper limit; \( T_{op} \) is the operative temperature of the zone and \( T_{upper} \) is the upper limit for 80% of acceptability, for the occupied hour \( i \). The upper limit temperature is based on the month’s outdoor mean temperatures. EHF is then calculated, according to Equation 4.

\[
EHF = \frac{\sum_{i=1}^{N} E_{\text{hot},i}}{N} \tag{4}
\]

Where \( N \) is the number of occupied hours.

Since Brazilian typical climates do not have extreme cold temperatures, occupants tend to adapt to cold temperatures by increasing clothing (clo) (De Vecchi et al., 2015). For this reason, the lower limits were not taken into account for the EHF.

**Sensitivity analysis**

The performance of ANNs depends on how the input data correlates to the outputs, as well on how the data is described to the inputs. For that, a Sensitivity Analysis (SA) was conducted to be able to identify the most influential parameters. Second order and total effects were also analyzed. While first order and total effects indices are obtained for each parameter, second order effects are analyzed for every combination between two parameters. Some inputs could have a minor influence on the outputs when analyzed separately, but interact with other inputs in a way that significantly impact the output, which is the reason to analyze the higher order effects.

SA was taken using the method of Sobol’ (1993). This method is applied on a sample created specially for the SA (Sobols sampling). For this analysis, 108 000 cases composed the sample. This process was performed with SALib (Herman and Usher, 2017), a Python library developed for SA procedures.

SA was conducted considering three different outputs: EHF; annual average Operative Temperature; and annual average Air Changes per Hour (ACH). A single simulation performed on Energyplus obtains these three outputs, so the only difference between the analyses is the output vector. The inputs are based on one single sample, which is developed for the SA method.

Based on the results of the SA, the sample used to train the ANN was generated varying only the most significant parameters. Non-influential parameters had their values fixed with their most common value found in the real buildings database.

**Surrogate model**

The dataset used for training was developed from 112 000 simulation results, sampled by Sobol’s sampling method. The validation was conducted using a sample of 30 000 cases generated by LHS. In the validation dataset, the parameters not included as inputs of the ANN were set equal to the fixed values of the training sample. The choice of using different sampling methods is to avoid the existence of any bias related to the sampling method.

Next, the dataset created for the SA was used to assess the performance of the ANN when the values of the parameters not considered as inputs vary. Thereby, the SA sample simulation results were available and the sample was different from the ones used for the ANN training and validation.

The ANN was developed using Python programming language, with the library TensorFlow (Abadi et al., 2015).

Definition of the hyperparameters is an iterative process. Different numbers of layers and nodes can obtain results with different performances. The optimization algorithm chosen for the learning process also influences the results. Another important concern is the processing of the data. Normalization, transformation and other processes applied on the dataset features before the training step should be observed. All these aspects were considered to define the final configuration of the ANN.

The performance of the ANN is measured by its mean absolute error, and the absolute error of the 95th percentile.

**Results**

**Base model**

On the validation of the single zone approach, the average difference between EHF of the thermal zones considering the whole building and single zone models was 0.015. This approach allowed the parametrization of all variables. Therefore, even though there are some errors related to the simplified model, it allows the development of a more accurate ANN.

**Sensitivity analysis**

Figures 6 and 7 present the results of Sobol’s sensitivity analysis for first order and total effects, related to EHF and annual average Air Changes per Hour (ACH). These indices are proportional to the influence of the input on the output.

Occupant density is the most influential parameter on EHF, since it obtained the highest index value for first order, second order and total effects. From the first order effects indices, it was observed that, after occupant density, the contact with the ground and the window opening factor are the most influential parameters on both EHF and average Operative Temperature. The index values for these two outputs are similar overall, since EHF is calculated from the Operative Temperature. Therefore, the annual average Operative Temperature results were not plotted. The most influential parameters on the ACH are, as expected, the ones related to the zone’s openings. They are: opening factor of the window; exposure condition of walls and windows; and WWR.
Significant colinearities are identified among the parameters. Even though the zone area has a higher first order index, the exposure of the walls and the roof could be more influential on the EHF, when combined with certain parameters. Second order effects analysis shows high interaction between the exterior walls exposure and the opening factor of the window. The interaction could be explained due to the the higher influence of either one or two windows on the facade when the window opening factor is higher. The highest second order effect index for EHF was the one relating occupant density and ground exposure. These two parameters were also the two most influential for first order and total effects analyses.

Thus, this strong relation points out how contact with the ground can help to dissipate internal heat gains at the climate of Sao Paulo.

Based on the SA results, some the least influential parameters were not considered as input features for the final model. They are: height of the zone’s floor; SHGC; wall thermal capacity; shading device; building ratio.

**Surrogate model**

From the 112 000 simulations generated to train the ANN, EHF results varied from 0.06 to 1.00, according to Figure 8.
Table 3: Parameters with fix values. Not considered by the surrogate model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Fix Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shading [m]</td>
<td>0</td>
</tr>
<tr>
<td>Wall Thermal Capacity [kJ/m²K]</td>
<td>155</td>
</tr>
<tr>
<td>SHGC [−]</td>
<td>0.87</td>
</tr>
<tr>
<td>Floor Height [m]</td>
<td>15</td>
</tr>
<tr>
<td>Building ratio [−]</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Roof and ground exposure were defined as a binary feature, corresponding to 0 (adiabatic) and 1 (exposed). One parameter was not represented as a number. Instead, it was transformed into a string feature. The string represented the zone’s walls and windows exposure. The rest of the parameters were normalized to values between -1 and 1.

The final model was an ANN with two hidden layers: one with 50 nodes; the other with 20 nodes. The chosen algorithm for the optimization of the model was the Adagrad’s Optimizer (available on TensorFlow), with a learning rate of 0.05.

Figure 9 presents a scatter plot comparing results between simulations and predictions for the validation dataset, which had the parameters not considered as inputs for the ANN training set with the same values as the ones fixed in the training sample. The mean absolute error was 0.008, and the absolute error of the 95th percentile was 0.024.

For the dataset used in the SA, the mean absolute error between simulation and prediction was 0.040, and the absolute error of the 95th percentile was 0.129 (Figure 10). This dataset did not have any of the inputs set on a fixed value. The ANN overestimates the EHF for most of the compared cases, a bias that occurs due to the choice of the fixed values. Therefore, some cases have lower EHF calculated by the simulation because of their shading device, or lower SHGC.

Conclusion

Natural ventilation is a passive cooling strategy for buildings. It works in a complex way, and its simulation can be expansive, specially in the first stages of the building design. A surrogate model could be of great use to estimate thermal comfort in early design phases in a simple and fast way. Thus, it simplifies the estimation of different scenarios with low computer cost. The performance results of the proposed ANN for the climate of Sao Paulo, Brazil, showed that it is possible to develop a simple model, with few input parameters, that could be useful for architects, designers and other decision makers in the early stages of building design.

Further studies could include characteristics of the climate as input features of the surrogate model. This would allow the development of a model suitable for different cities around the world.

Acknowledgment

The authors would like to thank the support of the National Council for Scientific and Technological Development - CNPq.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior Brasil (CAPES) Finance Code 001

References

ergonomic systems. Software available from tensorflow.org.


