# COMPONENT-BASED MACHINE LEARNING MODELLING APPROACH FOR DESIGN STAGE BUILDING ENERGY PREDICTION: WEATHER CONDITIONS AND SIZE

Sundaravelpandian Singaravel[1], Philipp Geyer[1], Johan Suykens[2]
[1]Architectural Engineering Division, KU Leuven, Belgium
[2]ESAT-STADIUS, KU Leuven, Belgium
sundar.singaravel@kuleuven.be

## Abstract

Building energy predictions are playing an important role in steering the design towards the required sustainability regulations. Time-consuming nature of detailed Building Energy Modelling (BEM) has introduced simplified BEM and metamodels within the design process. The paper further elaborates the limitations of this method and proposes a component-based Machine Learning Modelling (MLM) approach which could potentially overcome the current limitations.

The paper proposes a methodology for developing component-based MLM that generalise well. Generalisation, in this paper, refers to the reusability of an MLM developed with data from a specific situation in similar circumstances. As a first step in ongoing research on component-based MLM, a model is developed with data from a simple box building with weather data of Amsterdam, Brussels and Paris and two occupancy profiles. It is shown that the MLM is able to predict the annual energy for (1) same box building under different weather conditions not included in the training data (2) different dimensions of the box building for one case weather data and occupancy. The prediction error for annual heating demand is lower than 10% for all evaluated cases while the prediction error for annual cooling demand ranges -3.4% to 28.3%. Good generalisation is observed for all heating energy predictions whereas only for a few cooling energy predictions. Possibilities for model improvement and next steps of the research project are described.

## Introduction

Typically, Building Energy Models (BEM) evaluate the performance of a building design upon completion. Stringent sustainability requirements created a need for the use of BEM during early design stages. However, detailed BEM is time-consuming for early design, while simplified BEM could result in prediction gap (Singaravel & Geyer, 2016). Limitations of current simple BEM is; it typically focuses on a specific BEM area with simplification in other model areas. For example, simplified BEM to explore architectural elements usually have simplified HVAC (Heating, Ventilation and Air Conditioning) model (Miyamoto, et al., 2016). Resulting in limited cross-discipline interactions during early design. This is due to its time-consuming nature and lack of energy modelling experts at early design stage.

The need for having models with high accuracy and low computation time is increasing with our need to evaluate many design options at the early design stage and caused by the increasing complexity of the sustainable building. Metamodels developed with BEM results provide a flexible workflow; a simple input structure for obtaining curtailed information contained within a simulation model is ideal for early building design (Henry, et al., 2016). Metamodels also have high calculation speed suitable for early design (Van Gelder, et al., 2014). A simple method of this category is the Response Surface Method (Box & Draper, 2007). Such surrogate models were used to represent energy simulation results as well as to monitor real performance exceeding the thermal behaviour of buildings (Chlela, et al., 2009, Jaffal, et al., 2009, Catalina, et al., 2013, Geyer & Schlüter, 2014).

Machine Learning Models (MLMs) extends this potential with large and diverse datasets, which is not only valuable for building design but also for building stock management. Artificial Neural Networks (ANN) serves to model building performance, which is time-series prediction of energy consumption, in many studies (e.g., Neto & Fiorelli, 2008, Ekici & Aksoy, 2009, Gao, et al., 2010, Ahmed, et al., 2011, 2011a, Stavrakakis, et al., 2012, Kusiak & Xu, 2012, Catalina, et al., 2008, Naji, et al., 2016). Support Vector Regression (SVR)—another machine learning method— is also frequently used (e.g., Li, et al., 2009, de Wilde, et al., 2013, Jain, et al., 2014). Simpson, et al. (2001), Ashtiani, et al. (2014) and Wei, et al. (2015) compare methods of surrogate modelling, partly in the context of the built environment. Yang, et al. (2005) and Moon (2012) propose models that adapt during prediction. Furthermore, due to reduced computation times, metamodeling has been exploited for optimisation of buildings (e.g., Eisenhower, et al., 2012, Ekren & Ekren, 2008, Zhang, et al., 2012).

Another growing trend is the availability of wide variety of data, ranging from time series data (example: monitoring houses for the IEA EBC Annex 58, Strachan, et al., 2015) to point estimates from project databases of government or sustainability certification bodies like LEED (Leadership in Energy and Environmental Design). The available data is useful for a design decision. Current

machine learning/metamodelling approach limits the use of wide variety of data. Time series approaches include prediction on historical data. The time element present within these methods is similar to dynamic annual energy simulations. Typically, machine learning models developed today are using point estimate data such as annual energy consumption. The limitations of current approach are (1) the inability to quantify the contribution of a design element on the energy prediction and the support of respective engineering reasoning for improving the design and (2) incorporate available interesting time series data like the IEA EBC Annex 58 monitoring data (Strachan, et al., 2015) which could support design stages to achieve more accurate predictions.

To overcome the limitations presented above, a component-based approach for MLMs is proposed, which offers the following benefits compared to current MLM:

- Ability to quantify the contribution/effect of a design element on the energy prediction;
- Applicability in new situations of building design which opens the possibility of reusable MLM;
- Have a modular nature which is suitable for applying in the building design, especially linked to building information modelling (BIM).

This paper elaborates first findings in terms of the feasibility of a component-based MLM performance prediction in early design phases. For that purpose, a MLM for a simple box building is examined for its generalization in terms of weather conditions and different building dimensions. This experiment examines the feasibility of a method to elaborated on more complex situations in future and provides indication of feasibility but no complete proof-of-concept. The paper is structured in the following manner:

- Proposed method for development of component-based MLM
- Case-based development and evaluation of component-based MLM
- Discussions and conclusions

## Proposed method for development of component-based MLM

Based on several tests the method outlined in this section has been developed. This method is domain-neutral, which means that it can also be applied to other types of building performance simulation, such as daylighting analysis, where computation time is high.

Component-based MLM is developed through the following steps:

1. **Identification of the performance parameters** to be estimated. Example: Energy performance of a building design.
2. **Decomposition of a calculation methodology** to identify the required model structure.
3. **Data collection**. Data source can be simulations or monitoring data from sensors or statistical data or other data sources. Before using in the following steps, proper data cleaning and transformation must be applied.
4. **Input parameter (or feature) selection** using engineering knowledge, statistical and feature selection methods for effective generalisation and to observe all the required interactions.
5. **Train, cross-validate and test component MLM.** For the selected ML algorithm after training, cross-validation and testing are required to identify the need for more training data or input parameters or MLM tuning and to evaluate generalisation.
6. **Use of component-based MLM** to steer the design towards the requirements/objectives either interactively or supported by search algorithms.

## Case-based development and evaluation of component-based MLM

### Development of Component-Based MLM structure

The objective of this section is to identify a model structure for the component-based MLM approach that could highlight its potential. A simple building energy calculation method is decomposed to understand the parameters and energy flow within a calculation method. Equation 1 shows a formula used to estimate the heating load of a building.

$$Q_{\text{Heating}} = Q_{\text{t}} + Q_{\text{inf}} + Q_{\text{v}} - Q_{\text{sol}} - Q_{\text{occ}} - Q_{\text{apl}} \quad (1)$$

This equation indicates the basic model structure required to estimate the heating load of a building, which consist of the following components:

1. Losses through transmission ($Q_t$) and infiltration ($Q_{inf}$)
2. Ventilation load ($Q_v$)
3. Gains through solar irradiation ($Q_{sol}$), occupancy ($Q_{occ}$) and appliances ($Q_{apl}$).

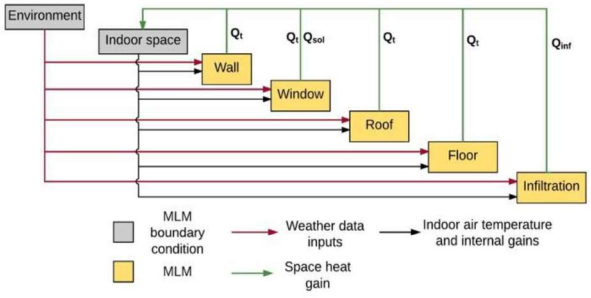


*Figure 1 Structure of component-based MLM*

Based on Equation 1, the component structure for MLM is developed (as shown in Figure 1), which estimates heat gains and losses through wall, floor, roof, windows and infiltration. Since the main objective of the study is to present a component-based MLM approach, ventilation

load and appliances gain are neglected and ideal HVAC efficiency is used for this study.

The outputs of sub-MLMs are used to aggregate the annual building heating load. Heating energy demand is estimated by adding the hourly heating load at indoor heating set-point. Similarly, an equivalent equation can be used to derive cooling load which is in turn used to estimate cooling energy. Note that the component MLM for the wall, floor, roof, windows and infiltration estimates both transient heat- gain and loss, which are used to aggregate the heating and cooling energy demands.

Input parameters for the components are selected based on physical equations in combination with domain knowledge on factors which influence heat gain and loss. Figure 2 shows the input and output for each MLM components.

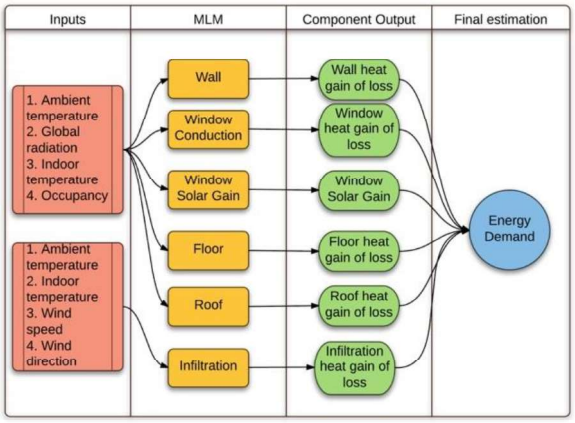


*Figure 2 Input and output structure within the component-based MLM*

In this paper, generalisation refers to model reusability in similar conditions to the data used in its development. This gives an indication on the expected performance of the model on similar unseen or new data. Further research is required to understand or standardise the conditions for evaluation of MLM generalisation. Since the main objective of the paper is to present a methodology for developing component-based MLM, the test case is limited to two situation.

Evaluation of MLM generalisation is performed by changing the model's location and dimensions. Hence, building properties shown in Table 1 remain constant within the study. Therefore, they are not included as model inputs. The inclusion of these parameters within the model should increase MLM's generalisation, which will be evaluated in future research and is not covered in this paper.

**Data collection**

Training data for the identified input and output parameters is obtained for a simple box building located in Amsterdam, Brussels, Paris *with-* and *without occupancy* using parametric BEM. 10% of the training data (selected randomly) is kept aside to evaluate and to refine the performance of MLM. This data is referred as cross-validation data.

Data is also collected from (1) the same building model located in London *with-* and *without occupancy, occupied between 8:00 to 18:00* (2) different dimensions of the box building located in Brussels with 100% occupancy. This data is used to test the generalisation of component-based MLM. In this paper, occupant behaviour is just the presence (100% occupancy) and absence (0% occupancy) of occupants; no other interactions are considered.

*Description of BEM model*

A simple box building model is developed in IES VE. The geometry of the building is based on BESTEST Case 600 (see Figure 3). Table 1 shows the properties of the building characteristics used within the energy model. In this model, ideal efficiencies of HVAC system are used, i.e., building heating and cooling energy demand is equal to building heating and cooling load. Furthermore, heating and cooling profiles were set to 'on continuously'.

*Table 1 Building characteristics*

| | Description | Properties |
|---|---|---|
| Wall | Area: 64.2 $m^2$ | U-Value: 0.26 $W/m^2K$ |
| Window | Area: 12 $m^2$ | U-Value: 1.6 $W/m^2K$ g-value: 0.4 |
| Floor | Area: 48.8 $m^2$ | U-Value: 0.22 $W/m^2K$ |
| Roof | Area: 48.8 $m^2$ | U-Value: 0.18 $W/m^2K$ |
| Occupancy | Density: 10 $m^2$/pp | Sensible: 90 W/pp Latent: 60 W/pp |
| Airtightness | 0.25 ach | |

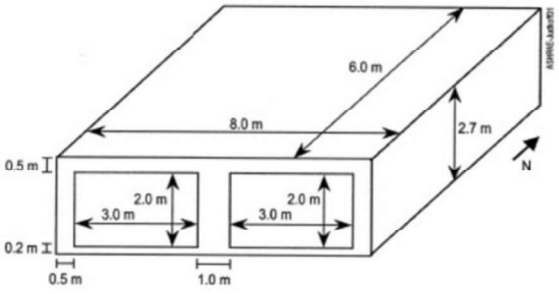


*Figure 3 Geometry of the simple box building (Hopfe, et al., 2007)*

*Description of training data*

Climate data obtained from Amsterdam, Brussels and Paris are used to train the MLMs. Figure 4 and Figure 5 show the frequency distribution of the weather and heat

gain data used to train and cross-validate the MLMs. Description of weather data is as following:

- Average dry-bulb temperature is 10.5°C with a maximum of 35°C and a minimum of -9.1°C
- Average global radiation is 113.5 W/m$^2$ with a maximum of 902.5 W/m$^2$. However, majority of the time a global radiation of 7 W/m$^2$ (inferred through the median) is observed
- Average wind speed is 4.6 m/s with a maximum of 22 m/s
- Predominant wind direction is between 210 to 240.

The indoor temperature and occupancy gains are also acquired for each simulation time step. Indoor temperature ranges between 20°C and 25°C, boundaries of the indoor temperature range also correspond to heating and cooling set points. Occupant gains alter between 0 and 100%.

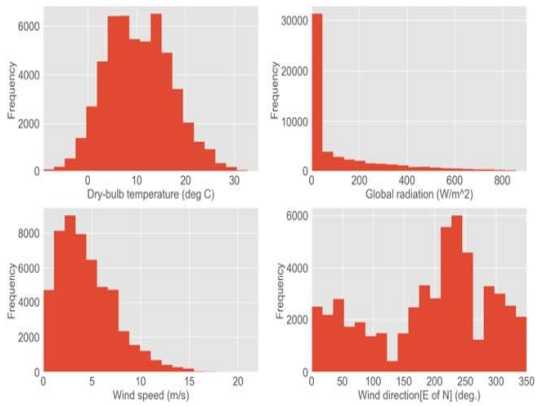


*Figure 4 Histogram of training weather data (inputs)*

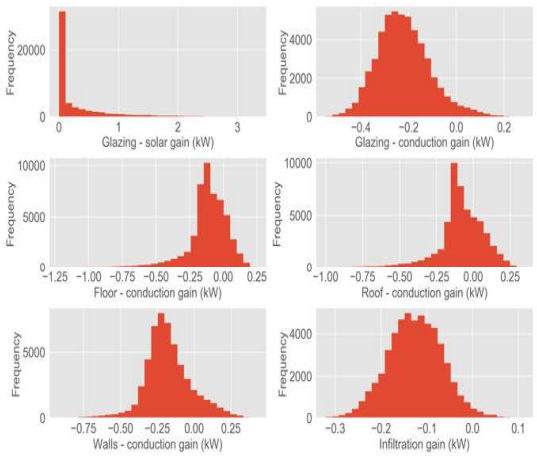


*Figure 5 Histogram of training heat gain and loss data (output)*

From Figure 5, the following can be noted:

- Training data predominantly consist of heat loss through the building envelope, and heat gain is less. However, the occurrence of extreme heat losses within the dataset is low.
- Heat conduction through the window, wall and infiltration gain are (approx.) normally distributed while the distribution is skewed for other building elements.

*Description of test data*

In this paper, training data is generated by varying weather data only. Hence, the model should generalise if the test data is within the distribution of the training data. Test data is generated from (1) London weather (*with- and without occupancy, occupied between 8:00 to 18:00)* and (2) different dimensions of the box model with Brussels weather (*always occupied*), to evaluate generalisation. For this paper, energy is predicted for different dimensions by scaling the MLM output values. Scaling is the process of converting MLM response value into values per meter square or cube and multiplying it with appropriate dimensions. For instance, infiltration gain is divided by volume of base dimension model and multiplied by volume of corresponding cases. Table 2 shows the dimensions used to test validity of model for cases ranging up to five times the floor area.

It is required that weather conditions and building envelope heat loss characteristics are similar to the training data set. Otherwise, resulting model performance will not be good. Furthermore, it is not required that the same combination of input and response values which occur in the test data be present in the training dataset.

*Table 2 Evaluated dimensions for model validation*

| | Base dimensions | Case 1 | Case 2 | Case 3 |
|---|---|---|---|---|
| **Volume (m$^3$)** | 131.7 | 191.5 | 343.6 | 652.8 |
| **Floor (m$^2$)** | 48.7 | 70.9 | 127.3 | 241.8 |
| **Roof (m$^2$)** | 48.7 | 70.9 | 127.3 | 241.8 |
| **External wall (m$^2$)** | 64.2 | 75.4 | 100.8 | 137.4 |
| **Window area (m$^2$)** | 12.0 | 17.0 | 24.0 | 36.0 |
| **Occ gain (kW)** | 0.4 | 0.6 | 1.1 | 2.2 |

## Selection of ML algorithm and features

Model variations can be obtained by modifying the model structure (example: neural network with 5 or 10 hidden units) or by using different ML algorithms. The performance of an algorithm or model structure depends on the dataset (Alpaydın, 2010). The selection of ML algorithm and final features/inputs is done through evaluation of coefficient of determination ($R^2$) on cross-validation data. The method for selecting of ML algorithm and features are as follows:

- Selecting ML algorithm which has the highest cross-validation $R^2$ for all or majority of the component dataset
- Selecting features/input parameters (if required), to improve MLM performance
- Tuning of hyperparameters, to further improve model's performance.

*Selection of ML algorithm*

The MLM used in this study are modeled in Python using scikit-learn library (Pedregosa, et al., 2011). The algorithms evaluated are briefly explained in this section,

followed by identification of the algorithm used to model components. The ML algorithms for regression assessed in this paper are:

- ***Random Forest (RF)*** developed based on the concept of regression tree which splits training data based on variable into a tree structure. Single trees are not sufficient to develop a good regression model. Hence a group of regression trees are developed for predictions (Ma & Cheng, 2016). The predictions of each tree model within RF algorithm is averaged based on the probability of the prediction, to obtain a final prediction (Pedregosa, et al., 2011). The hyperparameter (default) values used to train an RF model are:
  1. *Number of trees*: 10
  2. *Measure of quality*: Mean Squared Error
  3. *Max features*: Equal to number of features
  4. *Minimum sample split*: 2
  5. *Minimum sample leaf*: 1
  6. *Bootstrap*: True
- ***Extremely Randomized Trees (ERT)*** is developed based on an ensemble of regression trees. The main difference between ERT and other tree methods is that the splits node is chosen randomly and the entire training data is used for developing the tree (Geurts, et al., 2006). Hyperparameter (default) values used to train an ERT model are same as those of RF except *bootstrap* which is False.
- ***K-Nearest Neighbors (k-NN) regressor*** learning centers *k*-nearest neighbors for each examination point. *K*-most similar value located within training data and weight function are used for predictions. Hyperparameter (default) values used to train k-NN models are:
  1. *Number of neighbors*: 6
  2. *Weights*: Uniform
- ***Multi-Layer Perceptron (MLP)*** is a feedforward neural network with one or more hidden units. MLP can learn non-linear function approximates for a set of input and output values (Cigizoglu, 2004). Hyperparameters used to train MLP models are:
  1. *Number of hidden layer*: 1
  2. *Number of units in a hidden layer*: 50
  3. *Activation*: Rectified linear unit function

Table 3 shows component-wise $R^2$ for training and cross-validation data for each ML algorithm and MLM component. From this table, it can be observed that all ML algorithms perform similarly on cross-validation data. This may not be the situation once the quantity or the nature of data changes. Furthermore, it can be noted that RF and ERT have lower cross-validation $R^2$ compared to their training $R^2$. This indicates that overfitting of data is taking place. For this study, ML algorithm which has the highest overall $R^2$ on cross-validation data is used. RF has high cross-validation $R^2$ for the majority of the component dataset. Hence, all components are modelled with this algorithm.

*Table 3 Coefficient of determination (R2) with training and cross-validation data for different ML algorithm*

| Component MLM | Coefficient of Determination (R^2) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | RF | | ERT | | k-NN | | MLP | |
| | Training data | Cross-validation data | Training data | Cross-validation data | Training data | Cross-validation data | Training data | Cross-validation data |
| Wall | 0.9132 | **0.5593** | 0.9876 | **0.5333** | 0.6717 | **0.5526** | 0.5473 | **0.5496** |
| Window - Conduction | 0.9935 | **0.9741** | 0.9977 | **0.9716** | 0.9771 | **0.9693** | 0.9693 | **0.9702** |
| Window - Solar gain | 0.9757 | **0.8652** | 1.0000 | **0.8610** | 0.8705 | **0.8267** | 0.8050 | **0.8096** |
| Floor | 0.9166 | **0.7185** | 0.9589 | **0.7173** | 0.7740 | **0.7101** | 0.7020 | **0.7109** |
| Roof | 0.9083 | **0.7261** | 0.9458 | **0.7254** | 0.7826 | **0.7273** | 0.7227 | **0.7376** |
| Infiltration | 0.9983 | **0.9926** | 0.9995 | **0.9926** | 0.9899 | **0.9865** | 0.9915 | **0.9919** |

<u>*Selection of features/input parameters*</u>

Investigation on the reason for low cross-validation $R^2$ for the wall, window solar gain, floor and roof indicated that sufficient features or input parameters were not present to map all heat gains and losses accurately on new or unseen data. Hence, feature selection exercise is performed only for these components.

Additional inputs to the MLM components is selected by analyzing the correlation observed within the training data collected from Amsterdam, Brussels and Paris BEM. Figure 6 shows the correlation between independent and dependent parameters from the parametric BEM, clustered as heat map matrix. The heat map indicates the following:

- Solar azimuth has a strong correlation with conduction through the wall and, in contrast, a weak correlation with conduction through the floor and the roof and solar gain through window;
- Solar altitude has a strong correlation with conduction through floor and roof and with solar gain through the window and a weak correlation with conduction through the walls.

Since, correlation only indicates the presence of a linear relationship, while the presence of a non-linear relationship as well as causal relation could not be ruled out. Thus, a further engineering interpretation is required to select parameters. Because of this interpretation, both solar azimuth and solar altitude are incorporated within the models. Additional inputs/features have improved cross-validation $R^2$ for all the components (see Table 4) on an average, by 23% from the previous case.

*Table 4 Increase in R2 through feature selection and tuning of hyperparameter*

| Component MLM | ML algorithm | R^2 with Cross-validation data | | | Performance increase |
|---|---|---|---|---|---|
| | | Baseline | Feature select | Tuning | |
| Wall | RF | 0.5593 | 0.9424 | 0.9476 | **41.0%** |
| Window - Conduction | RF | 0.9741 | | 0.9755 | **0.1%** |
| Window - Solar gain | RF | 0.8652 | 0.9626 | 0.9655 | **10.4%** |
| Floor | RF | 0.7185 | 0.9236 | 0.9303 | **22.8%** |
| Roof | RF | 0.7261 | 0.9136 | 0.9229 | **21.3%** |
| Infiltration | RF | 0.9926 | | 0.9932 | **0.1%** |


*Figure 6 Correlation clustered heat map matrix for selecting inputs parameters for component MLM*

*Tuning of hyperparameters*

Turning of hyperparameters can improve MLM's performance further. Using validation curve method, the required number of trees for the RF algorithm is determined, which has further enhanced the performance by 1% (on an average, see Table 4). Minimum sample split is also identified for floor and roof MLM using the same method.

**Development of ML component model**

Selected ML algorithm is trained with data from *Amsterdam, Brussels* and *Paris*. The developed component-based MLM give a transient response and predicts heating and cooling load based on annual weather data. Annual heating and cooling energy demand are determined by adding heating and cooling loads at indoor temperature set points.

**Evaluation of generalisation**

Generalization is evaluated by predicting annual energy demand for the following cases:

*Same box building with London weather and three occupancies:*

Table 5 shows the estimated $R^2$ with data collected from London BEM. $R^2$ on test data ranges between 0.7155 and 0.9909, with an average of 0.8912.

*Table 5 Coefficient of determination (R2) on hourly predictions from MLM components and energy demand*

| Component MLM | Training data R^2 | Cross-validation data R^2 | Test data R^2 | | | |
|---|---|---|---|---|---|---|
| | | | No occupancy | Occupied between 8-18 | Always occupied | Average R^2 |
| **Wall** | 0.9919 | 0.9476 | **0.8681** | **0.8826** | **0.8407** | 0.8638 |
| **Window - Conduction** | 0.9946 | 0.9755 | **0.9545** | **0.9450** | **0.9380** | 0.9458 |
| **Window - Solar gain** | 0.9960 | 0.9655 | **0.9546** | **0.9541** | **0.9540** | 0.9543 |
| **Floor** | 0.9801 | 0.9303 | **0.8488** | **0.8353** | **0.8538** | 0.8460 |
| **Roof** | 0.9838 | 0.9229 | **0.8517** | **0.8273** | **0.8708** | 0.8499 |
| **Infiltration** | 0.9986 | 0.9932 | **0.9909** | **0.9894** | **0.9870** | 0.9891 |
| Hourly Heating Energy | N/A | | **0.9059** | **0.8968** | **0.8873** | 0.8967 |
| Hourly Cooling Energy | | | **0.7155** | **0.7256** | **0.9102** | 0.7838 |

Figures 7- 9 shows a section of hourly energy predictions for both component-based MLM and BEM with London data. It can be noted from these figures that the predictions from MLM follow the prediction trends of BEM. For heating, the hourly heat flows leading to the heat demand are covered very well. For cooling, the occurrence of cooling demand is also identified well whereas the cooling peak has a slight offset that leads to the deviation in prediction and needs to be corrected by a model improvement. The instability or fluctuations observed within the MLM predictions could be the results of no time decay attribute present within the MLM, i.e. the effect of a prediction at a time step is not observed in the following time step.

Table 6 shows the annual energy predicted from MLM and BEM. The prediction errors are lower than 8% in all the cases, except for annual cooling energy for the case with occupancy between 8 and 18.

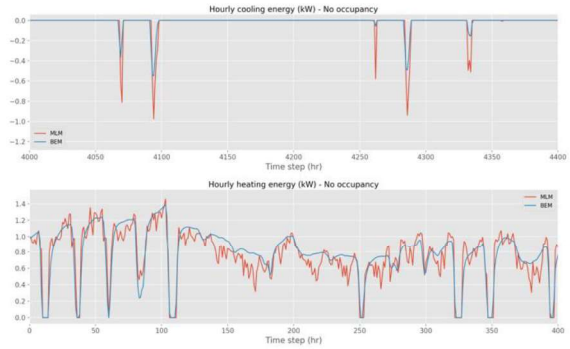


*Figure 7 Hourly energy prediction London data with no occupancy (orange- MLM prediction and blue- BEM prediction)*

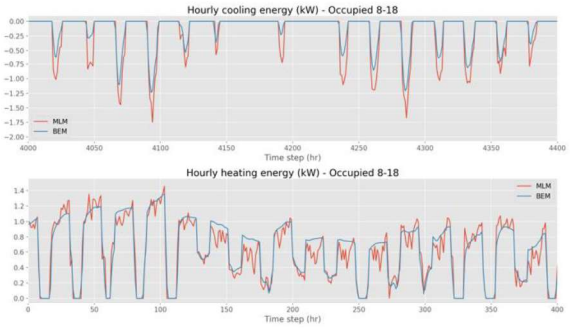


*Figure 8 Hourly energy prediction London data with 100% occupancy between 8 - 18 hours (orange- MLM prediction and blue- BEM prediction)*

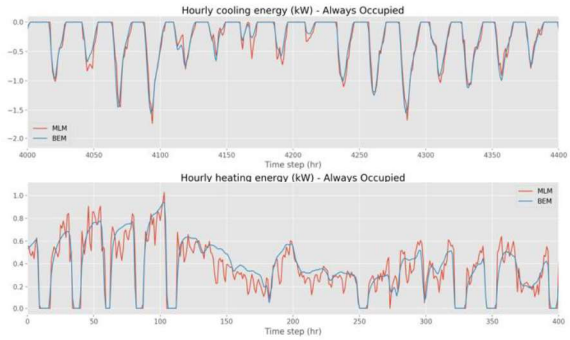


*Figure 9 Hourly energy prediction London data with 100% occupancy (orange- MLM prediction and blue- BEM prediction)*

*Table 6 Comparison of annual energy predictions from MLM and BEM*

| Case | MLM - Annual Heating Energy (kWh) | BEM - Annual Heating Energy (kWh) | Error | MLM - Annual Cooling Energy (kWh) | BEM - Annual Cooling Energy (kWh) | Error |
|---|---|---|---|---|---|---|
| No occupancy | 3624.36 | 3609.89 | **0.4%** | 92.11 | 85.06 | **7.7%** |
| Occupied between 8-18 | 2746.56 | 2664.67 | **3.0%** | 565.16 | 412.29 | **27.0%** |
| Always occupied | 1310.90 | 1368.97 | **-4.4%** | 838.28 | 866.47 | **-3.4%** |

*Different dimensions of the box model with Brussels weather and 100% occupancy:*

For a simple exemplary case transfer validation, dimensions of the box building MLM are changed and compared with the results of conventional BEM. Besides the base case, three further cases shown in Table 2 are tested. Figure 10 and 11 shows the annual energy predicted by MLM and BEM. MLM predictions are close to BEM predictions in all cases, except for case 3 cooling energy prediction. The reason is that the scaled effect of accumulated errors over the components are resulting in a lower cooling energy prediction than BEM prediction. This will improve as the goodness of fit increases with more training data and better feature selection.

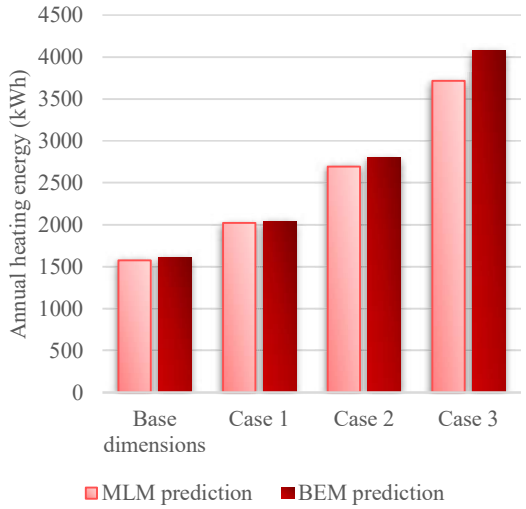


*Figure 10 Comparison of annual heating energy (kWh)*

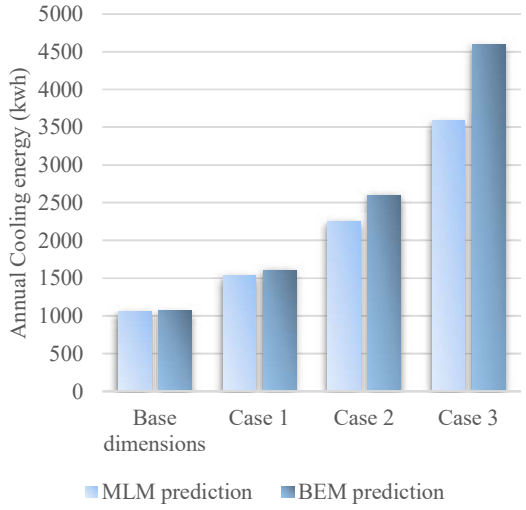


*Figure 11 Comparison of annual cooling energy (kWh)*

## Discussion

The paper shows a study of component-based machine learning for energy demand prediction. Exemplary results on weather conditions and on scaling indicate that the proposed method leads to a component-based MLM that generalizes well. We plan more testing for understanding the limits of generalization in more complex situations in future research. This research will apply the method to realistic building cases.

Good generalization, as shown for different weather conditions and size of the box building, is vital for design stage applications with its variation. Benefits of component-based approach compared to a monolithic approach are the ability to:

- Identify important components for an energy prediction and reduction;
- Quickly predict hourly energy demand of a building design, which is essential to capture dynamic behavior and peaks while looking at renewable energy strategies for the building design;
- Flexibly adapt/update a specific component instead of re-training the whole ML model in the case of single ML model for predictions.

The test case with occupancy between 8-18 has high error for cooling energy prediction compared to other cases (see Table 6). The high error is not the result of bad generalization. On the contrary, the high error is due to lack of physical interactions observed through thermal mass within the MLM. Thermal mass reduces the cooling demand predicted through BEM. Since, neither the training data nor the inputs/features captures this physical phenomenon, MLM cooling energy predictions are higher than BEM predictions (refer cooling energy predictions in Figure 8). Such interactions can be captured within the training data by collecting data from a diverse or representative set of occupancies through sampling techniques like Latin-hypercube.

Evaluating MLM for different dimensions of the box model through scaling of output values shows that MLM's can be used in situations/cases that does not resemble the training case (see Figures 10 – 11). It also highlights that in a component-based arrangement, accumulated error plays an important role, as the errors are amplified while scaling. Hence, MLM's with high cross-validation and test $R^2$ has to be developed.

Furthermore, uncertainty in the prediction that are a result of accumulated error can be mitigated by incorporating prediction interval within the prediction process. Predictions with long interval width can be evaluated further with detailed BEM, and the results can be viewed with caution.

Finally, selection of input structure must be based on both engineering knowledge and statistical methods, such as correlation or mutual information. Feature selection methods such as recursive feature elimination could also be used to identify suitable input parameters for obtaining high model accuracy. Use of such techniques combined with engineering knowledge is a more robust method for identifying suitable input parameters compared to relying on only one of the methods. The methodology for the use of both engineering and statistical methods for input

parameter selection in building energy prediction will be researched further.

## Conclusion

In this paper, component-based MLM developed for a box building trained with weather data from Amsterdam, Brussels and Paris has been evaluated on its prediction quality on new weather data and dimensions of building. The developed MLM generalised well for heating predictions in all cases and for some cooling prediction cases. This result indicates that developing MLMs with diverse datasets and appropriate input parameters could result in models that generalise well under different design situations provided that the new data match the distribution of the training data. Further research on generalisation of component-based MLM for building design with additional data and input parameters will be performed, to identify the full potential of such an approach.

Increasing need to design and develop buildings with high energy efficiency has increased the need for performing design space exploration right from the early design stages. This requires models which are quick and accurate. Machine learning gives the opportunity to predict energy performance based on data with high accuracy and low computation time. Component-based ML modelling approach takes this a step further which are the possibility to:

1. Quantify the reason for a design performance prediction, which is typically not possible for monolithic whole building MLM.
2. Link each MLM component to a BIM element making it possible to have an energy prediction instantaneously after all the required components are present within the BIM environment.
3. Introduce monitored data obtained from manufacturers or other buildings into design stages, potentially closing the performance gap. The key criteria for data collected are that it should cover the design space in a representative manner.
4. Extended for other building performance simulations, such as CFD, lighting or acoustic simulations can dramatically reduce high computational load without compromising accuracy.

We expect that this potential will assign component-based MLM a significant role in performance-based building design, especially, in early design phases.

## Acknowledgements

The research is funded by STG-14-00346 at KUL. The authors acknowledges support of ERC AdG A-DATADRIVE-B (290923), KUL: GOA/10/09 MaNet, CoE PFV/10/002 (OPTEC), BIL12/11T; FWO: G.0377.12, G.088114N, G0A4917N; IUAPP7/19 DYSCO.

## References

Ahmed, A., Korres, N., Ploennigs, J. & Elhadi, H., 2011. Mining building performance data for energy-efficient operation. *Advanced Engineering Informatics,* 25(2), p. 341–354.

Ahmed, A., Otreba, M., Korres, N. & Elhadi, H., 2011. Assessing the performance of naturally day-lit buildings using data mining. *Advanced Engineering Informatics,* 25(2), pp. 364-379.

Alpaydın, E., 2010. *Introduction to Machine Learning.* 2nd ed. s.l.:The MIT Press.

Ashtiani, A., Mirzaei, P. & Haghighat, F., 2014. Indoor thermal condition in urban heat island: Comparison of the artificial neural network and regression methods prediction. *Energy and Buildings,* 76(0), pp. 597-604.

Box, G. & Draper, N., 2007. *Response surfaces, mixtures, and ridge analyses.* s.l.:John Wiley & Sons.

Catalina, T., Iordache, V. & Caracaleanu, B., 2013. Multiple regression model for fast prediction of the heating energy demand. *Energy and buildings,* Volumen 57, pp. 302-312.

Catalina, T., Virgone, J. & Blanco, E., 2008. Development and validation of regression models to predict monthly heating demand for residential buildings. *Energy and Buildings,* 40(10), p. 1825–1832.

Chlela, F., Husaunndee, A., Inard, C. & Riederer, P., 2009. A new methodology for the design of low energy buildings. *Energy and Buildings,* 41(9), pp. 982-990.

Cigizoglu, H., 2004. Estimation and forecasting of daily suspended sediment data by multi-layer perceptrons. *Advances in Water Resources,* 27(2), p. 185–195.

de Wilde, P., Martinez-Ortiz, C., Pearson, D. & Beynon, I., 2013. Building simulation approaches for the training of automated data analysis tools in building energy management. *Advanced Engineering Informatics,* 27(4), pp. 457-465.

Eisenhower, B., O'Neill, Z., Narayanan, S. & Fonoberov, V., 2012. A methodology for meta-model based optimization in building energy models. *Energy and Buildings,* 47(0), pp. 292-301.

Ekici, B. & Aksoy, U., 2009. Prediction of building energy consumption by using artificial neural networks. *Advances in Engineering Software,* 40(5), pp. 356-362.

Ekren, O. & Ekren, B., 2008. Size optimization of a PV/wind hybrid energy conversion system with battery storage using response surface methodology. *Applied Energy,* 85(11), pp. 1086-1101.

Gao, Y., Tumwesigye, E., Cahill, B. & Menzel, K., 2010. *Using data mining in optimisation of building energy consumption and thermal comfort management.* s.l., In Software Engineering and Data Mining (SEDM), 2010 2nd International Conference on. IEEE, pp. 434-439.

Geurts, P., Ernst, D. & Wehenkel, L., 2006. Extremely randomized trees. *Machine Learning,* 63(1), pp. 3-42.

Geyer, P. & Schlüter, A., 2014. Automated metamodel generation for Design Space Exploration and decision-making – A novel method supporting performance-oriented building design and retrofitting. *Applied Energy,* Volumen 119, pp. 537-556.

Henry, H., Katherine, F., Brian, B. & Nicholas, L., 2016. *Achieving Actionable Results from Available Inputs: Metamodels Take Building Energy Simulations One Step Further.* Pacific Grove, CA, ACEEE.

Hopfe, C. y otros, 2007. *Uncertainty Analysis For Building Performance Simulation – A Comparison Of Four Tools.* Beijing, IBPSA, pp. 1383-1388.

Jaffal, I., Inard, C. & Ghiaus, C., 2009. Fast method to predict building heating demand based on the design of experiments. *Energy and buildings,* 41(6), pp. 669-677.

Jain, R., Smith, K., Culligan, P. & Taylor, J., 2014. Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy. *Applied Energy,* Volumen 123, pp. 168-178.

Kusiak, A. & Xu, G., 2012. Modeling and optimization of HVAC systems using a dynamic neural network. *Energy,* 42(1), p. 241–250.

Li, Q., Meng, Q., Cai, J. & Yoshino, H., 2009. Applying support vector machine to predict hourly cooling load in the building. *Applied Energy,* 86(10), p. 2249–2256.

Liu, Y., Huang, Y. & Stouffs, R., 2015. Using a data-driven approach to support the design of energy-efficient buildings. *Journal of Information Technology in Construction (ITcon),* Volumen 20, pp. 80-96.

Ma, J. & Cheng, J. C., 2016. Identifying the influential features on the regional energy use intensity of residential buildings based on Random Forests. *Applied Energy,* Volumen 183, p. 193–201.

Miyamoto, A., Trigaux, D., Nguyen, T. V. & Troyer, F. D., 2016. *From a Simple Tool for Energy Efficient Design in the Early Design Phase to Dynamic Simulations in a Later Design Stage.* Zurich, Conference: Sustainable Built Environment (SBE) Regional Conference.

Moon, J., 2012. Performance of ANN-based predictive and adaptive thermal-control methods for disturbances in and around residential buildings. *Building and Environment,* Volumen 48, pp. 15-26.

Naji, S. y otros, 2016. Estimating building energy consumption using extreme learning machine method. *Energy,* Volumen 97, pp. 506-516.

Neto, A. & Fiorelli, F., 2008. Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption. *Energy and buildings,* 40(12), pp. 2169-2176.

Pedregosa, F. y otros, 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research,* Volumen 12, pp. 2825-2830.

Ritter, F., Geyer, P. & Borrmann, A., 2015. *Simulation-based Decision-making in Early Design Stages.* Eindhoven, The Netherlands, 32nd CIB W78 Conference.

Simpson, T., Poplinski, J., Koch, P. & Allen, J., 2001. Metamodels for Computer-based Engineering Design: Survey and recommendations. *Engineering with Computers,* 17(2), pp. 129-150.

Singaravel, S. & Geyer, P., 2016. *Simplifying Building Energy Performance Models to support an Integrated Design workflow.* Krakow, EG-ICE 2016.

Stavrakakis, G., Zervas, P., Sarimveis, H. & Markatos, N., 2012. Optimization of window-openings design for thermal comfort in naturally ventilated buildings. *Applied Mathematical Modelling,* 36(1), p. 193–211.

Strachan, P., Monari, F., Kersken, M. & Heusler, I., 2015. IEA Annex 58: Full-scale Empirical Validation of Detailed Thermal Simulation Programs. *Energy Procedia,* Volumen 78, pp. 3288-3293.

Van Gelder, L., Das, P., Janssen, H. & Roels, S., 2014. Comparative study of metamodelling techniques in building energy simulation: Guidelines for practitioners. *Simulation Modelling Practice and Theory,* October, Volumen 49, pp. 245-257.

Wei, L., Tian, W., Silva, E. & Choudhary, R., 2015. Comparative Study on Machine Learning for Urban Building Energy Analysis. *Procedia Engineering,* Volumen 121, pp. 285-292.

Yang, J., Rivard, H. & Zmeureanu, R., 2005. On-line building energy prediction using adaptive artificial neural networks. *Energy and Buildings,* 37(12), pp. 1250-1259.

Zhang, J., Chowdhury, S., Messac, A. & Castillo, L., 2012. A Response Surface-Based Cost Model for Wind Farm Design. *Energy Policy,* 42(0), pp. 538-550.