Cluster-Based Surrogate Modelling Approach to Remote Building Characterization

Shane Ferreira[1], Burak Gunay[1], Brent Huchuk[2], Scott Shillinglaw [2], Farzeen Rizvi [2]
[1]Carleton University, Ottawa, Canada
[2]National Resources Council, Ottawa, Canada

Abstract

To better inform retrofit decision-making and benchmark performance of the existing building stock, a scalable and computationally efficient method needs to be developed to estimate energy-related features. The conventional field-scale and inverse model-based characterization approaches are limited in their scalability due to high engineering costs and the need for high-fidelity building automation system (BAS) data. To this end, a cluster analysis-based method is developed to estimate the thermophysical envelope, HVAC operation, and casual gains features based on only heating and cooling load profiles. Firstly, variants of a mid-rise office building energy model were generated through random sampling of these features. Then, for each variant, three-parameter univariate heating and cooling change-point models were fitted separately for operating and afterhours. Different clustering techniques were applied to the change point model parameters of all variants. Building feature sets that are likely associated with each cluster are then identified. The method could effectively identify energy-related building features from simple heating and cooling load profiles.

Introduction

Retrofit decision-making is a time and labour-intensive process that requires in-depth analysis of the building features that impact energy performance. Determining energy-related building features typically requires ASHRAE Level 3 type energy audits, followed by performance assessments and diagnostics using building energy simulation tools to identify and evaluate the potential impact of retrofit opportunities (Ma et al., 2012). This process is limited in scalability as each analysis is unique to a specific building and is therefore not viable for wide-scale portfolio-based analysis of the building stock. A key challenge in portfolio-based retrofit decision-making lies in developing a low-cost method that is scalable to remotely characterize energy-related envelope, casual gains, and HVAC operation features without requiring a calibrated energy model for each building. This would allow for preliminary screening of the building stock and provide a basis to prioritize a building's retrofit or maintenance potential.

Conventional field-scale testing methods to determine thermal transmittance of the envelope (e.g., heat-flux method) tend to be labour intensive, intrusive to occupants, and present measurement uncertainties that arise from environmental conditions such as radiation and thermal mass effects (Rasooli and Itard, 2019). As such they are rarely employed in practice. In-situ methods to determine air permeability (e.g., fan pressurization test) are sensitive to stack and wind-induced pressures as they interrupt the fan's ability to maintain a constant pressure differential (Younes et al., 2012). The method also presents practical challenges in preparing large buildings for the test as intentional openings need to be closed, and the building needs to be vacated during testing. Both field-scale testing examples discussed above are inappropriate for wide-scale portfolio-based analysis as they require tightly regulated experiments to be carried out under strict conditions for successful application. Moreover, energy-related features are not limited to only envelope features but are also defined to a large extent by casual gains and HVAC operational features.

Inverse model-based methods allow building systems to be described remotely by considering how the system's inputs (e.g., measured energy use, CO_2 concentration) are mapped to the system's outputs (e.g., thermal transmittance, air permeance). These methods require access to BAS data to map the functional relationship between the system's inputs and outputs. With the recent increase in the distribution of temperature-based sensors and smart meters in buildings, large amounts of raw data have become available to be leveraged for remote characterization of building features. The accuracy of these models is dependent on the inputs used, data resolution and duration, and the model's ability to account for casual/solar heat gains (Gunay et al., 2021). To that end, it is also not uncommon for energy analysts to encounter gaps, stagnant values, or outliers in audited BAS data. For example, Ramallo-González et al., (2018) developed lumped parameter models, an inversed-based grey-box model, to determine the reliability of inverse modelling for wide-scale characterization of thermal properties using a minimum number of sensors. The study concluded that low order models were able to accurately estimate heat transfer coefficients and internal temperatures if the input sensor data from heating, electricity, infiltration, and ventilation was available and limited to the winter season only echoing the high-fidelity data limitations discussed above. Moreover, many existing buildings do not have the required sensors available to make this approach feasible for wide-scale portfolio-based analysis.

Calibrated energy simulation models, a type of white-box inverse model, can alternatively be used to characterize energy-related features. The calibration method utilizes physics-based building simulation tools (e.g., EnergyPlus) to iteratively tune the influential energy-related parameters, typically via a metaheuristic search algorithm, until an adequate agreement is reached with measured energy use data (Chong et al., 2021). Calibrated energy models carry high computational and engineering costs and are dependent on energy audits and other meta-data specific to the building being studied (WWR, orientation, gross floor area, etc.) to execute the process and is therefore not scalable to other buildings.

In brief, on-site characterization approaches are labour-intensive and costly, and should therefore be reserved for the most critical buildings. The need for low-cost remote characterization of energy-related features is partially addressed by inverse grey-box model-based approaches, albeit only for buildings with high-quality BAS data. Inverse white-box models can also be used in the characterization of energy-related features; however, they are associated with a high engineering cost and limited in scalability. To that end, surrogate models represent an untapped opportunity to complement existing on-site and remote characterization techniques for energy-related building features by negating many of their limitations.

A surrogate model is built from energy simulation input and output data gathered over several evaluations of an objective function (Westermann and Evins, 2019). The idea is to emulate high-fidelity energy models using a statistical model, thereby improving computational efficiency. Using surrogate models for characterization objectives is advantageous as it can reduce the high computational and engineering costs associated with developing calibrated energy models, the model can be developed remotely without requiring high-fidelity BAS data and is scalable to portfolio-based analysis of the building stock. Despite this suitability, the majority of the research incorporating surrogate models in building performance simulation addresses early design (Edwards et al., 2017), sensitivity analysis (Rivalinet al., 2018), or optimization (Wong et al., 2010), with only one notable study focusing on the remote characterization (Ascione et al., 2017). The study proposed artificial neural network (ANN) to predict dynamic energy performance, thermal comfort, and retrofit scenarios for a building category. In contrast, this study proposes a cluster-based surrogate approach to infer a building's energy-related features from heating and cooling load profiles.

Methodology

As shown in Figure 1, the methodology employed in this paper trains a surrogate model to associate change-point models with the most likely energy-related features. The datasets used for training the surrogate are thermophysical envelope, casual gains, and HVAC operation features used as inputs to EnergyPlus and the corresponding change point model outputs. Piecewise linear regression is performed on the heating and cooling load output to obtain change point models which are then clustered and compared to 35 surveyed building's change point models in Ottawa.

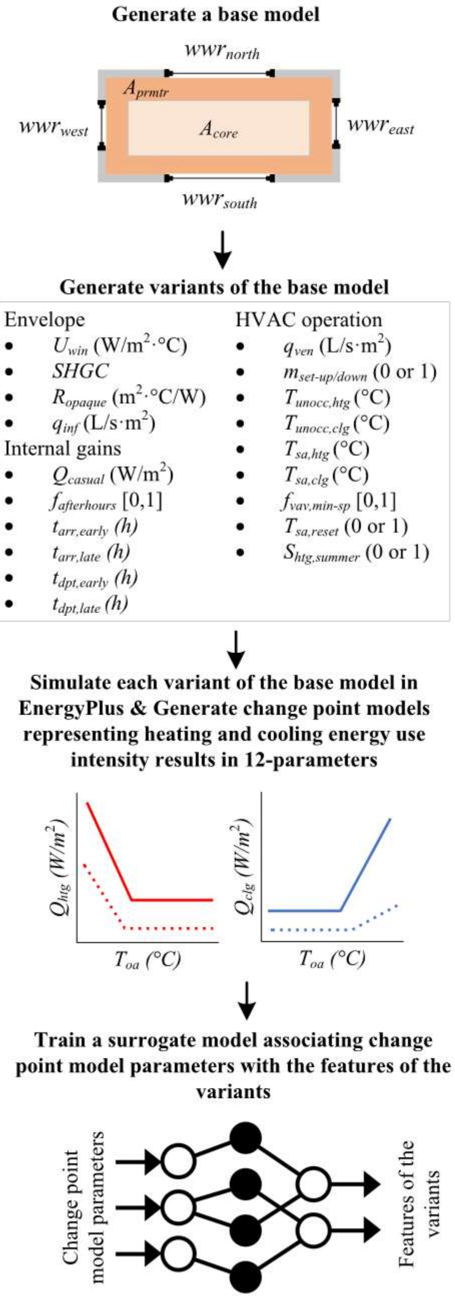

Figure 1: Methodology overview

Developing a calibrated base-model archetype

The base model by (Hobson et al., 2021) is selected to serve as an archetypical mid-rise office building. In this case, a six-story (8170 m^2) square office building with a 65% window-to-wall ratio (WWR) situated on Carleton University Campus in Ottawa, Canada (Climate zone 6A). It is connected to a local district steam heating and chilled-water cooling plant. The base model geometry in Figure 2 was created in SketchUp Make (2017) and imported into EnergyPlus V9.5.0 via Euclid plug-in. The EnergyPlus weather input file is an Ottawa-based 2019 Actual Meteorological Year (AMY) weather file. Solar distribution is set to full interior and exterior on all surfaces. An adaptive surface convection algorithm is used for interior and exterior surfaces. The zoning plan separates core zones (3550 m^2) from perimeter zones (4620 m^2). The internal gains, occupancy, electrical and lighting loads are all implemented homogeneously as variables based on zone area. Infiltration through the envelope assumes a constant flow rate through exterior surface areas.

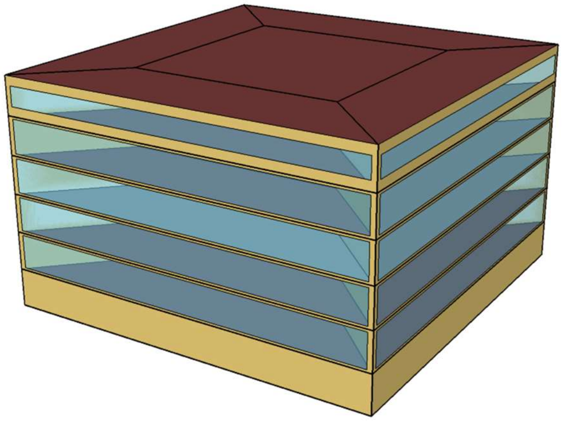

Figure 2: Overview of base model archetype geometry

The two air handling units (AHUs), 30 variable air volume (VAV) terminals, and 24 perimeter hydronic units are based on EnergyPlus template objects and configured with general BAS information. AHU operating schedules are from 04:45 am to 17:30 on weekdays only. The heating and cooling zone temperature setpoints are 22°C and 23.5°C during AHU operating hours, respectively. During operating hours, the VAV AHUs maintain the interior environment through heating and cooling coil use and perimeter hydronic units. Outside scheduled operating hours, the heating and cooling zone temperature setpoints are 20°C and 25°C, respectively. The VAV AHUs wake up outside scheduled operating hours if the unoccupied heating setpoint temperature is violated in any of the zones to provide supplementary heating to the perimeter hydronic units. Similarly, in the cooling season, if the cooling setpoint temperature is violated, the VAV AHUs wake up to provide cooling. A differential dry-bulb temperature-based economizer is available with a high limit of 18°C. Supply air temperature setpoint reset is programmed as a function of the outdoor air temperature.

Once the base model is built with the required metadata, EnergyPlus is coupled with MATLAB through a custom script to calibrate the base model via an optimization algorithm. The calibration process employs a metaheuristic search algorithm namely the genetic algorithm (GA) to tune unknown envelope, internal gains, and HVAC operation parameters to 2019 measured heating and cooling load data. A calibrated base model was only needed to assess the ability of the surrogate model to predict the energy-related features of the building archetype.

The coefficient of variation of the root mean squared error (CVRMSE) is used as the cost function to measure the difference between simulated and measured heating and cooling energy until prescribed calibration metrics are reached. ASHRAE Guideline 14 requires CVRMSE and Normalized Mean Bias Error (NMBE) less than 30% and 10% respectively for hourly whole building energy calibration (ASHRAE 2014).

The genetic algorithm is an optimization method that is based on a natural selection process. The algorithm repeatedly modifies a population of individual solutions by cross-mating the elite members to converge on an optimal solution. The model hyperparameters, listed in table 2, were selected such that the GA output results between several optimization runs varied negligibly. This provides a measure of stability to the calibrated thermophysical feature estimates. Ultimately, this required a population size of 50 over ten generations, with a 50% cross-over fraction of the population not including the elite members.

Generating variants of the base model

Table 2 presents the 19 selected building features and sampling ranges used to develop the surrogate model. The 19 features are sampled using a static sampling method, Latin hypercube sampling (LHS), where the number of returned samples is set to 1000. Static sampling implies that the surrogate model derivation happens sequentially – i.e., all sample locations are predefined before model fitting (Westermann and Evins, 2019). The feature set defined for each sample is used as inputs to the EnergyPlus base model archetype in successive simulation runs to generate a database of thermophysical input parameters and the corresponding heating and cooling load outputs. The objective in each simulation run is to extract univariate change point models from the heating and cooling energy data during HVAC operating and afterhours.

The thermophysical features parameter range, used to generate variants of the base model, is varied to be representative of a wide variety of operational inefficiencies found in practice. To that end, some variants do not have a supply air temperature setpoint reset programmed and instead have constant supply air temperature that ranges between 12°C - 15°C for cooling and 12°C - 24°C for heating. The minimum outdoor airflow fraction to the zone is also varied between 10% – 60%, and some variants have heating available in the cooling season.

Table 2: Selected input features for the surrogate model and sample range. The reference values that are parameter outputs from the calibration process are indicated with (*), the unmarked reference values are the default values included in the EnergyPlus model but were not included as optimization parameters during the calibration process.

Parameter Class	Parameter Description		Units	Ref. Value	Sampling Range
Envelope	$*U_{win}$	*Window u-value*	W/m²·°C	3	1.5 - 3.6
	$*SHGC$	*Solar heat gain coefficient*	~	0.3077	0.3 - 0.7
	$*R_{opaque}$	*Opaque thermal transmittance*	m²·°C/W	2.24	1.5 - 5
	$*q_{infil}$	*Infiltration rate*	L/s·m²	0.2651	0.1 - 1
Internal Gains	$*Q_{casual}$	*Light & plug load*	W/m²	15.66	3 - 15
	$*f_{afterhours}$	*Fraction on after hours*	%	0.75	0 - 1
	$*t_{arr,early}$	*First arrival*	h	8:36	05:00 - 11:00
	$*t_{arr,late}$	*Last arrival*	h	11:55	11:00 - 13:00
	$*t_{dpt,early}$	*Early departure*	h	12:00	13:00 - 17:00
	$*t_{dpt,late}$	*Last departure*	h	16:30	16:00 - 22:00
HVAC Operation	$*q_{ven}$	*Ventilation rate*	L/s·m²	0.8642	0.2 - 2
	$m_{set-up/down}$	*Night cycle*	~	1	0 or 1
	$T_{SA,clg}$	*Supply air temperature cooling*	°C	12	12 - 15
	$T_{SA,htg}$	*Supply air temperature heating*	°C	18	12 - 24
	$T_{sa,reset}$	*Supply air reset scheme*	°C	1	0 or 1
	$T_{unocc,htg}$	*Unoccupied thermostat setpoint heating*	°C	20	15 - 22
	$T_{unocc,clg}$	*Unoccupied thermostat setpoint cooling*	°C	25	25 - 30
	$f_{vav,min-sp}$	*Minimum outdoor air fraction*	%	0.2	0.1 - 0.6
	$S_{htg,summer}$	*Summer heating availability*	~	0	0 or 1

Univariate Change-point models

The univariate piecewise linear regression models in Figure 3 provides a fit between the heating and cooling energy use during each operating condition listed in table 3 and the outdoor air temperature. Each change point model provides three degrees of freedom (i.e., y-intercept, balance temperature, and slope) for each of the four-operating conditions resulting in a total of twelve change point parameters of interest that define each variant's energy use behaviour.

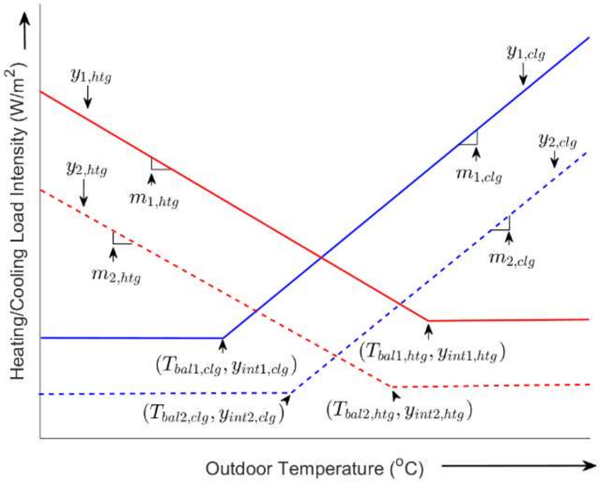
Heating/Cooling Load Intensity (W/m²)
$y_{1,htg}$
$m_{1,htg}$
$y_{2,htg}$
$m_{2,htg}$
$y_{1,clg}$
$m_{1,clg}$
$y_{2,clg}$
$m_{2,clg}$
$(T_{bal1,clg}, y_{int1,clg})$
$(T_{bal1,htg}, y_{int1,htg})$
$(T_{bal2,clg}, y_{int2,clg})$
$(T_{bal2,htg}, y_{int2,htg})$
Outdoor Temperature (°C)

Figure 3: A schematic illustrating univariate steady-state three-parameter heating and cooling change-point models during HVAC operational and afterhours conditions

Table 3: Description of the change-point models

Model	Description
$y_{1,htg}$	Heating energy use during operational hours
$y_{2,htg}$	Heating energy use during afterhours
$y_{1,clg}$	Cooling energy during operational hours
$y_{2,clg}$	Cooling energy use during afterhours

The y-intercept values (y_{int}) are a measure of the base-level heating or cooling energy use independent of outdoor air temperature. The slope values (m) indicate the rate of change in heating or cooling energy in response to outdoor air temperatures. The balance temperature indicates the outdoor temperature at which the model switches from weather dependent to weather independent behaviour. The balance temperatures ($T_{bal,htg}$) indicate the maximum outdoor air temperature below which the heating energy maintains a linear relationship with outdoor temperature. Likewise, values ($T_{bal,,clg}$) indicate the minimum outdoor air temperature above which the cooling energy maintains a linear relationship with outdoor temperature. The superscript (-) in equations (1) and (2) indicates only negative values of the expression are considered. Likewise, the superscript (+) in equations (3) and (4) indicates that only positive values are considered

$$y_{1,htg} = y_{int1,htg} + m_{1,htg}\,(T_{out} - T_{bal1,htg})^{-} \tag{1}$$

$$y_{2,htg} = y_{int2,htg} + m_{2,htg}\,(T_{out} - T_{bal2,htg})^{-} \tag{2}$$

$$y_{1,clg} = y_{int1,clg} + m_{1,clg}\,(T_{out} - T_{bal1,clg})^{+} \tag{3}$$

$$y_{2,clg} = y_{int2,clg} + m_{2,clg}\,\left(T_{out} - T_{bal2,clg}\right)^{+} \tag{4}$$

Cluster analysis

Clustering is an unsupervised machine learning technique used to find commonalities in a large dataset and group them based on shared characteristics. In the context of this study, the clustering algorithms find commonalities in 12 change point model parameters and groups the most likely building features associated with each change point cluster. These clusters can then be compared to surveyed buildings' change point model parameters to optimize the range of parameters needed as inputs to the surrogate model.

Before initiating the clustering algorithms, the change point parameter dataset is normalized to enable the comparison of the parameters. This is done by rescaling the range of values between zero and one but retaining the shape of the distribution. The features are then reduced through principal component analysis (PCA) to a subset of features that explain 95% of the total variance.

In all clustering techniques described below the number of clusters are predefined and determined by the Calinski-Harabasz Index (CH-Index) (Caliñski and Harabasz, 1974). The CH-Index assesses the optimal number of clusters in a dataset by formulating a metric that is defined as the ratio of the sum of between-cluster dispersion and inter-cluster dispersion for all possible clusters within a given range. The metric is higher when clusters are dense and well separated. In this study, the criterion's cluster range was 10 to 20 clusters.

The clustering procedure employs three clustering techniques and selects the best method based on the method's ability to partition the features into distinct groups as defined by the CH-Index. These three are the k-means, Gaussian mixture distribution, and hierarchical clustering techniques. Among them, k-means is a centroid-based clustering algorithm that iteratively seeks to minimize the squared Euclidean distance between all data points and perceived cluster centroids. Gaussian mixture distribution is a distribution-based clustering algorithm that seeks to fit a pre-defined number of probability density functions to the dataset. Hierarchical clustering is a connectivity-based clustering algorithm that separates the dataset by using a linkage matrix to define the dissimilarity of each data point from the rest of the dataset. The number of clusters and algorithm that maximizes the CH-Index is then selected for the analysis. To identify the building features that may generate each change point model cluster's profile, k-means clustering is employed on a normalized building feature dataset using the optimal clustering solution generated from the change point model clusters.

To validate the method's potential to predict a building's energy-related features, the calibrated change point model was compared to the clustered change point models and the cluster with the closest fit was identified. The predicted thermophysical, HVAC operation and internal heat gain features of that cluster are then compared to the calibrated thermophysical features.

Results and discussion

Figure 5 provides a statistical summary of the change point models' parameters over four operating conditions. The blue box is representative of the 25th and 75th percentiles and the red line represents the median. The whiskers of the boxplot enclose max and min values and the data points identified beyond the whiskers are considered outliers. Outliers are defined as data points exceeding 1.5 times the interquartile range away from the 25th or 75th quantiles. The filled data points are the calibrated change point model parameters relative to the variant's distribution.

The balance temperature indicates the outdoor temperature at which the model switches from weather-independent to weather-dependent energy use behaviour. The distribution of balance temperatures in Figure 5(A) could be influenced by a combination of the envelope's thermal performance, internal gains, and HVAC controls during each operating condition. In a case study analysis of 35 federal office buildings in Ottawa, Canada, Gunay et al. (2019) determined most buildings have a balance temperature less than 18°C with the median cooling balance temperatures (4°C) during occupied periods being significantly less than the median heating balance temperatures (18.5°C). This implies that the vast majority of surveyed governmental office buildings in Ottawa would heat and cool under identical conditions during occupied periods.

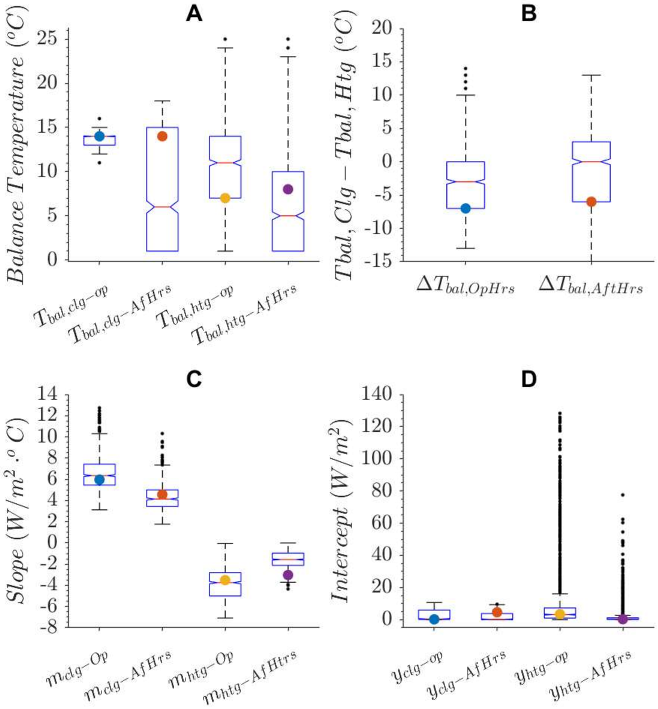

Figure 5: Variants' and calibrated model's change point parameter (filled scatter points) distribution for four operating conditions – heating and cooling energy use during HVAC operating and afterhours: (a) balance temperatures, (b) difference in balance temperature, (c) slope, (d) y-intercepts at balance temperature.

Figure 5(B) shows the difference between cooling and heating balance temperatures during operational hours ($\Delta T_{bal,1}$) and afterhours ($\Delta T_{bal,2}$). The distribution results indicate that 23% of the variants heat and cool simultaneously during operating hours and 35% afterhours. For operating hours, the 230 variants that heat and cool simultaneously are exclusively defined by inactive supply air reset schemes. During afterhours, if night cycling is inactive, the cooling balance temperatures are very low (1°C) due to the air system not cycling on zones that violate the afterhours setpoint temperature, resulting in those variants simultaneous heating and cooling.

Figure 5(C) describes the distribution of slope parameters - i.e., weather-dependent rate of change in heating or cooling energy use per degree change in outdoor air temperature. High cooling slopes and low heating slopes are indicative of poor-performing envelopes (Fels, 1986). The maximum cooling slopes determined by Gunay et al. in the case study was 5 W/m^2·°C and the minimum heating slopes were -5 W/m^2·°C. Heating slope values for both operating ($m_{1,htg}$) and afterhours ($m_{2,htg}$) less than -5 W/m^2·°C in the distribution are variants that could be considered over-ventilated. In contrast, cooling slope values greater than 5 W/m^2·°C in both operating and afterhours are with high ventilation rates, SHGC, lighting-plug densities, and heating supply air temperatures. Outlier cooling slope values (> 10 W/m^2·°C) in addition are correlated with inactive supply air temperature reset schemes and heating available in the cooling season. The minimum cooling slopes of the variants is 3 W/m^2·°C and 1.7 W/m^2·°C for operating and afterhours respectively. A majority of the government buildings studied by Gunay et al. (2019) had cooling slopes less than the minimum in the variant's distribution.

Figure 5(D) shows the distribution of the y-intercept values. The intercept values are defined as the minimum expected heating or cooling energy independent of exterior conditions. Concerning operational heating ($y_{int1,htg}$) and afterhours heating ($y_{int1,htg}$), outliers above 16 W/m^2 are correlated with high ventilation rates, high heating supply air setpoints, and heating available in the summer. The operational heating quantiles contain intercepts from 1 W/m^2 to 7.3 W/m^2 with a median value of 3.5 W/m^2. This is in line with the y-intercept distribution for occupied periods observed in the government buildings studied by Gunay et al. (2019). However, the government building's occupied cooling intercept distribution contains outlier cooling intercepts greater than the operational cooling intercepts ($y_{int1,clg}$) found in the variants' distribution.

In brief, the parameter distribution in most cases has good agreement with those reported by Gunay et al. (2019), but to overcome the slope and y-intercept distribution limitations mentioned above, the number of returned samples would have to be increased to allow for wider parameter distributions.

K-means clustering results

Figure 6 presents cluster analysis results of the variant's change point model compared to the calibrated change point model parameters used for validation. In this case, the CH-Index determined the optimal number of clusters to be 10 using the k-means algorithm. Cluster two most resembles the calibrated change point parameters. It is worth mentioning that the matching cluster is based on which method is used to determine the closest fit. In this case, the Euclidean distance was used.

Cluster two contains 70 variants that do not heat and cool simultaneously. The heating y-intercepts exceed the 75th quantiles but are not outliers while the cooling y-intercepts are contained within the 75th quantile. The heating slopes are contained within the 25th quantiles while the cooling slopes are contained within the 75th quantile. Clusters three (largest: 200 variants), four (140 variants), five (50 variants), eight (50 variants), and ten (150 variants) similarly do not heat and cool simultaneously but are differentiated based on slope and y-intercept parameters. Cluster three has similar slope characteristics to cluster two but has lower heating and cooling baseline energy use. Cluster four has identical heating slopes to cluster two but lower cooling slopes. Cluster four also has significantly lower heating intercepts than cluster two but greater cooling intercepts. Cluster five has lower heating slopes but identical cooling slopes. However, cluster five has significantly greater baseline heating (outlier at 44 W/m^2 for operating hours) and cooling energy use. Cluster eight and cluster five both have outlier baseline heating energy during operating hours at 44 W/m^2 and 81 W/m^2 respectively.

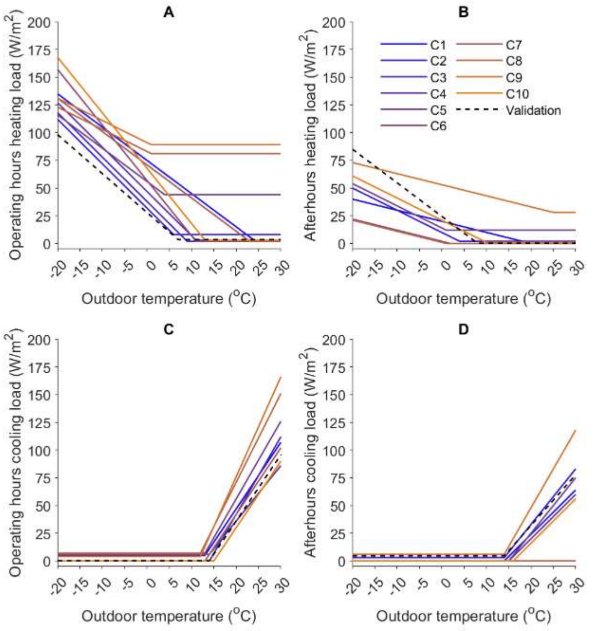

Figure 6: Variants' change point parameter clusters measured against calibrated model's change point parameters.

Cluster ten has lower y-intercepts and heating and cooling slopes than cluster two. Cluster six (160 variants) and nine (smallest: 20 variants) share the characteristic that they both simultaneously heat and cool in afterhours conditions but differ in slope and baseline energy use. Cluster six has smaller heating and cooling slopes. Cluster nine has an excessively high heating baseline (outlier at 81 W/m^2) and a higher operational cooling baseline than cluster 6.

Cluster one contains 70 variants and is the only cluster group that heats and cools simultaneously during both operating conditions. Cluster one is further characterized by heating and cooling slopes less than cluster two. Cluster one has a significantly smaller heating baseline but a greater cooling baseline when compared to cluster two.

Cluster seven contains 90 variants and is the only cluster that heats and cools simultaneously during regular scheduled operating hours. It has a heating slope characteristic comparable to cluster one and baseline characteristics comparable to cluster three.

Table 5 is the result of applying the optimal clustering solution from the change point parameters to the thermo-physical parameter dataset. It identifies a set of three common building features that are most likely to result in each cluster's change-point model. In this case, cluster two subset one features most resemble the calibrated model's features.

From the envelope parameter class results, we can deduce the method underestimates window U-Value and overestimates SHGC, wall R-Value, and infiltration rate.

None of the lighting-plug density cluster centers agree well with the calibrated model's casual heat gains parameters. It is worth noting that the highest casual gain cluster center is identified and there is good agreement between predicted and calibrated afterhours fractions.

The HVAC operation parameters agree fairly well with the calibrated model features but underestimate the zone level ventilation rates and overestimate the fraction of outdoor air to zones. There is also a rather large discrepancy in the unoccupied thermostat setpoint during the cooling season and heating availability in the summer.

The above-mentioned discrepancy could be rectified by increasing the internal gains parameter range and the number of returned samples to 10,000 to capture all the variations in the solution space. This will also more than likely result in a larger number of optimal clusters improving the accuracy of the parameter estimates.

Table 5: Clustered building feature look-up table - features based on optimal change-point model clustering solution

Cluster	Envelope				Casual		HVAC Operation								
	U_{win}	*SHGC*	R_{opaque}	q_{infil}	Q_{casual}	$f_{afterhours}$	q_{ven}	$m_{set-up/down}$	$T_{SA,clg}$	$T_{SA,htg}$	$T_{sa,reset}$	$T_{unocc,htg}$	$T_{unocc,clg}$	$f_{vav,min-sp}$	$S_{htg,summer}$
	W/m^2·°C	~	m^2·°C/W	L/s·m^2	W/m^2	%	L/s·m^2	[0,1]	°C	°C	[0,1]	°C	°C	%	[0,1]
Calibrated Model	3.0	0.31	2.2	0.265	15.7	0.75	0.864	1	12	18	1	20	25	0.2	0
C1	2.2	0.54	2.8	0.642	7.5	0.51	0.601	1	14	15	0	21	26	0.5	0
	2.7	0.46	4.0	0.439	7.1	0.78	1.076	1	14	21	0	18	27	0.3	0
	2.6	0.58	3.0	0.639	11.6	0.52	0.737	1	13	20	0	19	27	0.3	0
C2	2.7	0.46	3.2	0.492	11.9	0.72	0.705	1	13	17	1	19	28	0.4	1
	2.5	0.59	3.6	0.504	10.9	0.70	0.642	1	13	15	1	18	27	0.5	0
	2.4	0.48	3.4	0.520	9.9	0.47	1.222	1	13	18	0	18	28	0.4	1
C3	2.3	0.55	3.4	0.724	11.6	0.41	0.691	0	14	15	1	17	27	0.4	0
	2.5	0.46	2.8	0.397	8.6	0.62	1.407	0	14	17	1	20	28	0.3	0
	2.5	0.49	3.3	0.477	9.7	0.60	1.094	0	13	17	1	18	28	0.3	1
C4	2.9	0.51	3.6	0.459	7.5	0.50	1.064	1	13	18	1	18	27	0.3	0
	2.4	0.51	2.9	0.595	8.4	0.41	1.371	1	14	14	0	18	27	0.3	0
	2.3	0.48	3.5	0.576	9.3	0.61	1.004	1	13	20	1	18	27	0.3	1
C5	2.5	0.53	2.5	0.701	10.4	0.68	0.383	1	13	14	0	18	29	0.4	0
	3.1	0.36	2.9	0.383	7.0	0.70	1.384	1	13	21	0	18	27	0.4	1
	2.4	0.47	3.0	0.382	11.6	0.24	1.339	1	14	19	0	20	29	0.3	1
C6	2.8	0.53	2.9	0.734	7.6	0.43	1.087	0	14	15	0	18	27	0.4	1
	2.6	0.49	3.0	0.617	8.9	0.41	1.371	0	13	18	1	18	28	0.3	0
	2.6	0.48	3.4	0.516	8.3	0.41	1.078	0	14	18	1	19	27	0.3	1
C7	2.9	0.49	3.3	0.681	7.7	0.37	0.802	0	13	19	0	21	26	0.4	0
	2.6	0.53	3.1	0.564	7.5	0.31	1.135	0	14	21	0	17	28	0.5	0
	2.3	0.50	3.4	0.551	11.0	0.76	0.951	0	13	20	0	18	29	0.3	0
C8	2.6	0.61	4.3	0.719	10.8	0.25	1.173	0	13	22	0	19	28	0.4	1
	2.2	0.45	2.7	0.594	7.2	0.67	0.819	0	13	21	0	20	27	0.3	1
	2.8	0.52	2.8	0.391	9.0	0.53	1.327	0	13	20	0	17	27	0.3	1
C9	2.2	0.61	3.9	0.523	10.3	0.77	0.770	1	13	22	0	20	27	0.2	1
	1.9	0.53	4.0	0.697	11.5	0.47	0.863	1	13	22	0	16	29	0.4	1
	2.9	0.61	4.1	0.193	8.3	0.42	1.413	1	13	20	0	17	27	0.5	1
C10	2.7	0.47	2.9	0.651	7.2	0.29	1.281	1	14	16	0	19	28	0.3	0
	2.7	0.43	3.2	0.612	7.5	0.53	1.319	1	14	18	1	18	27	0.3	1
	2.5	0.50	3.1	0.528	8.3	0.38	1.590	1	14	18	1	19	27	0.3	0

Conclusion and future work

This paper has presented a cluster analysis-based surrogate modelling method to characterize thermophysical features of the envelope, HVAC operation, and casual gains based on heating and cooling load profiles for a particular archetype – a mid-rise, square-profiled office. We envision this method to form the basis of a screening and characterization tool for large-scale, portfolio-based remote characterization projects.

First, a calibrated base model is generated to serve as a building archetype. The energy-related building features are selected and sampled using a static sampling method. The building features are then fed as inputs into a building energy simulation tool. Then univariate change point parameters are extracted from the simulated heating and cooling load for HVAC operating and afterhours conditions to build a surrogate model database. The change point parameters are then clustered, and the optimal clustering solution is applied to the building energy-related features dataset.

The preliminary results presented in this paper are promising and are worthwhile to pursue further. The method could effectively identify building features based on the clustering solution but at this stage is far from being perfect. The method's identified limitations include inadequate distribution range, particularly concerning slope and intercept values and inadequate internal gains hyperparameter range. To rectify the limitations of the method and increase the method's accuracy, a larger number of samples and clusters need to be generated so that the distributions contain all possible change point parameters. Beyond addressing the limitations discussed above, planned future work includes expanding the sampling strategy to include geometric parameters that vary the WWR and aspect ratio between the core and perimeter zoning. It is also worth exploring whether increasing the change point model degrees of freedom will enable better clusters and more accurate results concerning HVAC-related energy use features. We also plan to extend the unsupervised machine learning technique used in this paper to include multiple inputs multiple output neural networks.

Acknowledgements

This research is supported by a research contract with the National Research Council Canada.

References

Ascione, F., N. Bianco, C. de Stasio, G.M. Mauro and G.P. Vanoli (2017). Artificial neural networks to predict energy performance and retrofit scenarios for any member of a building category: A novel approach. *Energy 118*,999–1017.

ASHRAE. (2014). ASHRAE Guideline 14: Measurement of Energy, Demand, and Water Savings.

Caliñski, T. and J. Harabasz (1974). A Dendrite Method Foe Cluster Analysis. *Communications in Statistics 3*,1–27.

Chong, A. and Y. Gu, and H. Jia (2021). Calibrating building energy simulation models: A review of the basics to guide future work. *Energy and Buildings 253*,

Edwards, R. E., J. New, L.E. Parker, B. Cui, and J. Dong (2017). Constructing large scale surrogate models from big data and artificial intelligence. *Applied Energy 202*,685–699.

Fels, M. F. (1986). PRISM: An Introduction. *Energy and Buildings 9*,5-18

Gunay, H.B., D. Darwazeh, S. Shillinglaw, and I. Wilton (2021). Remote characterization of envelope performance through inverse modelling with building automation system data. *Energy and Buildings 240.*

Gunay, H.B., W. Shen, G. Newsham, and A. Ashouri (2019). Detection and interpretation of anomalies in building energy use through inverse modeling. *Science and Technology for the Built Environment 25*,488–503.

Hobson, B. W., & Abuimara, T., & Gunay, H.B., & Newsham, G.R., (2021). How do buildings adapt to changing occupancy? A natural experiment. *Proceedings from eSim 2021: Building Simulation Conference*. Victoria (Canada), 14– 16 June

Ma, Z., P. Cooper, D. Daly, and L. Ledo (2012). Existing building retrofits: Methodology and state-of-the-art. *Energy and Buildings 55*, 889-902

Ramallo-González, A. P., M. Brown, E. Gabe-Thomas, T. Lovett, and D.A Coley (2018). The reliability of inverse modelling for the wide scale characterization of the thermal properties of buildings. *Journal of Building Performance Simulation 11*,65–83.

Rasooli, A. and Itard, L. (2018). In-situ characterization of walls' thermal resistance: An extension to the ISO 9869 standard method. *Energy and Buildings 179*,374–383.

Rivalin, L., P. Stabat, D. Marchio, M. Caciolo, and F. Hopquin (2018). A comparison of methods for uncertainty and sensitivity analysis applied to the energy performance of new commercial buildings. *Energy and Buildings 166*,489–504.

Westermann, P. and R. Evins (2019). Surrogate modelling for sustainable building design – A review. *Energy and Buildings* 198,170–186.

Wong, S. L., K.K.W. Wan, and T.N.T Lam (2010). Artificial neural networks for energy analysis of office buildings with daylighting. *Applied Energy 87*,551–557.

Younes, C., C.A. Shdid, and G. Bitsuamlak, (2012). Air infiltration through building envelopes: A review. *Journal of Building Physics 35(3),*267 – 3