Investigating Thermal Performance of Residential Buildings in Cold Climates

Aya Doma

A Thesis

In the Department

of

Building, Civil, and Environmental Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of

Master of Applied Science in (Building Engineering) at

Concordia University

Montreal, Quebec, Canada

January 2021

Concordia University

School of Graduate Studies

This is to certify that the thesis prepared

By: Aya Doma

Entitled: Investigating the Thermal Performance of Residential Buildings in the Cold Climates
Using a Smart Thermostat Dataset

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Building Engineering)

Signed by the final Examining Committee:

________________________________Chair
Dr. U. Eicker

______________________________Examiner
Dr. U. Eicker

______________________________Examiner
Dr. C. Lai

______________________________Supervisor
Dr. M. Ouf

Approved by____
Dr. Michelle Nokken, Graduate Program Director

January 11, 2021 _______________________________________
Dr. Mourad Debbabi, Dean, Gina Cody School of Engineering and Computer Science

Abstract

Investigating Thermal Performance of Residential Buildings in Cold Climates

Aya Doma

At least 65% of the existing residential building stock will still be in use by 2050, thus retrofitting existing buildings will be critical to reduce energy consumption. Prioritizing these retrofits typically requires thorough evaluation of the envelope's thermal performance, and the traditional methods to undergo such evaluation (e.g. energy audits) can be cost prohibitive, especially if it aims to cover hundreds or thousands of buildings. To this end, this study presents a novel data-driven approach to investigate the thermal performance of existing buildings using data collected from smart thermostats. The study focused on more than 60,000 houses across North America and relied on real-time indoor and outdoor temperature measurements at 5-minute intervals over a period of four years. Two grey-box modelling approaches namely, least-squares fitting of 1) decay curves, and 2) numerically integrated thermal energy balance equations were used to estimate a thermal time constant for each house. This time constant represented the time it takes for a house to achieve a new thermal equilibrium in response to changes in its internal and external thermal conditions. The resulting time constant values from both models were used to estimate lower and upper bound effective R-values for the entire envelope of each house. These results were also analysed with respect to ASHRAE climate zones, building-age, building-style, and floor-area. Finally, a classification model was developed to identify the time constant range for houses based on their attributes. The classification model indicated that floor area and ASHRAE climate zone were the most influential factors on time constant values obtained using both methods. By using a large sample size covering thousands of buildings nationwide, results of this research can be used to prioritize retrofits for existing buildings and can provide inputs for urban-scale energy simulations.

To my supportive parents: Mona Badr and Abdelhalim Doma

Acknowledgements

First and foremost, I would like to express my deepest appreciation to my supervisor Dr. Mohamed Ouf for his continuous support, patient guidance, and encouragement. I am very grateful for the opportunity to work under his supervision and for his substantial amount of effort and guidance during the period of this research.

I would also like to thank National Research Council Canada (NRC) for supporting and providing the necessary information for this research.

My grateful thanks are also extended to my examiners for their time in reviewing my thesis and for their constructive comments.

I would like to acknowledge my indebtedness and render my warmest thanks to my beloved parents, Dr. Mona Badr and Dr. Abdelhalim Doma, and my siblings, Dr. Hebatullah and Dr. Mohamed Doma, for everything I was, am, and will be. I appreciate their unconditional love, prayers, sacrifices and support at every stage of my life.

Last but not least, I must also acknowledge the support and encouragement I got from my colleague Eng. Ahmed Zalook and my research group in Intelligent Buildings and Cities lab (IBCL) at Concordia university.

Contents

List of Figures

List of Tables

List of Abbreviations

U-value	Thermal Transmittance
AMI	Advanced Metering Infrastructure
IOT	Internet of Things
RC	Thermal Time Constant for the entire building
R	Thermal Resistance
ARX	AutoRegressive models with eXogenous inputs
C	Thermal Capacitance for the entire building
ε	Energy loss rate
RK	Resistance- Heating Power averaged for entire building
DYD	Donate Your Data dataset
T_{in}	Indoor Temperature measurements
T_{ext}	Outdoor Temperature measurements
dt	Time step
$\dot{Q}_{in}$	The Internal Heat Gain
$\dot{Q}_{h}$	Heat flow supplied by heating system
$\dot{Q}_{sol}$	The solar radiation gains.
$\dot{Q}_{ven}$	Heat flow due to ventilation
$\boldsymbol{\delta_{on}}$	Duty cycle for the heating system
K	Heating Power
T_o	Initial indoor temperature
t	Time elapse

WCSS	With respect to the within cluster sum of squares
X_k	Cluster centroid
d	Datapoint in the cluster
ML	Machine learning algorithm
AdaBoost	Adaptive boosting algorithm
W_i	Initial weight for datapoints in AdaBoost
SMOTE	Synthetic minority oversampling technique
KNN	K nearest neighbour
RF	Random forest algorithm
KS	Kolmogorov–Smirnov test
R^2	R-squared
$\bar{T}$	mean of the indoor temperature reading.
$\widehat{T_{f\imath t}}$	the fitted indoor temperature.
T_{act}	the actual indoor temperature reading.

Chapter 1: Introduction

1.1 Background and Motivation

Recent projections anticipated 50% increase in the global energy consumption by 2050, 65% of which accounts for the residential and commercial buildings' energy consumption [1]. Moreover, the united nations projects that by 2050, at least 65% of the current residential building stock will still be in use [2]. Therefore, the need to put in action strategies to design energy-efficient buildings and prioritize retrofits for existing buildings is essential. Analyzing the thermal performance of existing buildings and identifying the attributes that influence such performance are key factors in identifying targeted retrofits in existing buildings. Typically, such analysis requires detailed knowledge of the building components' thermal properties such as their thermal resistance values which is identified using energy audits that may include onsite measurements. For example, Biddulph et al. used heat flux meters and thermistor temperature sensor to collect data from 93 occupied residential building in England. The collected data were then used to quantify the correlation between the heat flux and the difference between indoor and outdoor temperatures and estimate the building envelop thermal transmittance (U-value) [3]. Moreover, Aznar et al. collected data using surface temperature sensors at different layers of the building envelope within a residential building in Spain. This data along with indoor and outdoor temperature measurements was then used to train a deep learning model to describe the thermal behavior of this building [4]. Despite the accuracy of these techniques, they are time and cost intensive which is why they are typically limited to a number of buildings with sample sizes that may not be representative of the building stock.

With the increased adoption of advanced metering infrastructure (AMI) and the Internet of

Things (IoT) including smart thermostats in buildings, new opportunities arise to use data-driven approaches to investigate the thermal performance of existing buildings at a much larger scale. Unlike energy audits or onsite measurements, these datasets include measurements of indoor and outdoor conditions of a wide range of buildings which allows for a significantly quicker analysis of the thermal behaviour of large numbers of buildings. Furthermore, these datasets include other information about building attributes such as building age, building size, location, the number of occupants etc. These attributes can be then used to group buildings into different clusters and analyze their thermal behaviour in more detail.

1.2 Problem statement

Developing an effective policy to retrofit the existing building stock requires detailed information on the thermal performance of existing buildings. To this end, data-driven approaches can be used to estimate building thermal properties using smart thermostat data at a large scale (i.e., covering tens of thousands of buildings). Although different approaches have been introduced to analyze smart thermostat datasets, research on applying these approaches to derive indicators of buildings' thermal performance remains limited. Furthermore, the relationship between building physical attributes and their thermal performance at such scale has rarely been investigated. Such analysis can be beneficial for code officials and policymakers to confirm the effect of previous code changes on the building stock and can inform the development of new energy codes and retrofit programs. Results can also be used to represent the thermal properties of the existing building stock in urban scale energy simulations to investigate the effect of different retrofit scenarios on urban and regional energy consumption.

1.3 Scope and Objectives

The goal of this research is to investigate data-driven approaches to characterize the thermal performance of residential buildings in the colder climate (i.e. ASHRAE climate zone 4 to 8) in North America in relation to their attributes. The analysis relies on smart thermostat readings at 5-minute intervals from over 60,000 residential buildings across Canada and the United States for a period of up to four years, with the following specific objectives:

i. Estimate a time constant (RC) for each house, which represents the time it takes for a building to achieve a new thermal equilibrium in response to changes in its internal and/or external environmental conditions. Two different methods are investigated to identify the best method to estimate reliable results for more houses.

ii. Use the obtained RC values to estimate the corresponding R-value range for each house.

iii. Investigate the relationship between the estimated RC values and building's attributes such as age, style, area, and climate zone, and ranking the importance of these attributes on the identified RC values.

iv. Develop a classification model to predict the RC ranges for houses based on their attributes.

1.4 Thesis Organization

This research is composed of five chapters as follows. Chapter 1 provides the introduction and a summary of this thesis objectives. Chapter 2 contains a review of the literature to identify the main approaches used for whole-building thermal energy modelling, as well as the different approaches for estimating the thermal properties for existing buildings. Chapter 3 provides a description of the dataset used in this analysis as well as the detailed methods used to achieve the research objectives. Chapter 4 provides the results of this analysis as well as well as the performance evaluation of the prediction models. Finally, Chapter 5 provides the summary and the conclusions, lessons learnt and recommendations for future research.

Chapter 2: Literature Review

This chapter provides a review of the different approaches used to model the thermal energy in buildings and the advantages and challenges for each approach. This is followed by describing the different techniques used to estimate the thermal properties of existing buildings, especially focusing on those that can be applied to smart thermostat data.

2.1 Building Thermal Energy Modelling

Building thermal energy models are used to predict the thermal response of a building. Based on the objective and the goal of each model, the whole building can be modelled, or the models can focus only on the critical components of the buildings system such as the photovoltaic systems [5], solar thermal systems [6-8], heat pumps [9], air handling units [10], radiant floor slabs [9], boilers [11,12], etc. This section focuses on whole building thermal energy modelling using two main approaches, the forward modelling approach (white-box models) and data-driven modelling approach.

2.1.1 Forward Modelling Approach (White-box Models)

The forward modelling approach or the white-box modelling approach uses detailed physics-based equations to model building components (e.g. walls, windows, roof, etc.) and systems (e.g. HVAC system, lighting system, etc.) to predict the whole buildings dynamic thermal behaviour, such as the energy consumption and indoor comfort [13].

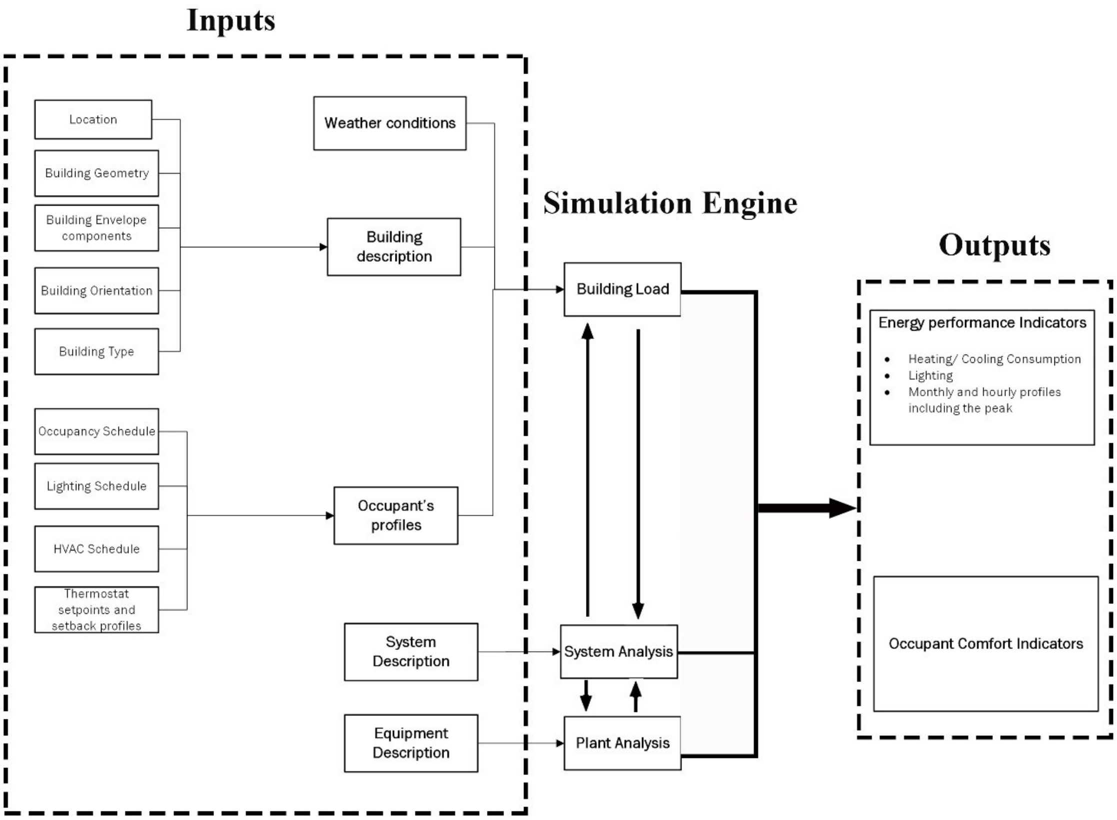


Figure 1: The general outline of the White-box model procedure and the main steps to detailed simulation [14].

The general outline for white box modelling is summarized in Figure 1. In this approach, inputs should first be identified then fed into the simulation engine which is a group of mathematical and physics equations that simulate the building operation and calculate the building energy consumption [14]. The inputs of the model can be categorized in five groups, the parameters for weather condition, the building description, the occupant's profiles, the system description and the equipment description. The weather conditions include the dry and wet bulb temperatures of outdoor air, solar radiation intensity, wind speed, etc. The building description mainly include data describe the location, design, construction materials, thermal zones, and infiltration. The occupant's profiles include the occupancy schedule and level of activity, usage profiles, internal heat gains, lighting and HVAC schedules etc. For the building system description, the system types and sizes, as well as the requirements for each system are required. Finally, the building equipment description addresses the HVAC components including the equipment types and sizes, performance characteristics, load assignments and auxiliary equipment [15]. Many of the building energy simulation software are based on the previous outline of the forward energy modelling (e.g., EnergyPlus, TRNSYS, DOE-2, ESP-R, etc.) [13].

2.1.2 Data-driven Modeling Approach (Black-box Models)

Instead of relying on detailed physics equations which require a high-level of knowledge about the building thermal properties, data-driven approaches use statistical methods to model the thermal behaviour of the buildings with no information about the building thermal properties.

Black-box models need data over a certain period of time to train the models to be able to predict the building operation under different conditions. Neural networks [16], [17- 19], time series models such as AutoRegressive models with eXogenous inputs (ARX) [20-22], and machine learning models [23, 24] such as linear regression, random forest and supportive linear regression models have been used to predict buildings' energy consumption. For example, Xu et al. trained artificial neural network model to predict multi-building energy use. To train their model, they used three years of monthly energy use data from seventeen buildings in China. The selected buildings covered four types of buildings, office buildings, educational buildings, laboratory buildings and residential buildings [19]. On the other hand, Kontokosta and Tull used dataset from 20,000 buildings in New York city to train machine learning predictive models including linear regression, random forest and support vector regression to predict building-level energy usage and energy use intensity for different building types (e.g. residential, industrial, etc.) [23].

Black box models are easy to build and computationally efficient, however, they require a large training data with wide range of operational conditions to avoid forecasting errors [13]. Moreover, they have poor generalization capabilities if the training dataset did not cover different types of buildings at different conditions.

2.1.3 Hybrid Modeling Approach (Gray-box models)

In some cases, detailed white-box modelling inputs that describe building envelopes and building systems may not be available, especially at a large scale in the context of urban-scale energy simulation [25]. Even if these inputs are available, the computational power needed for detailed white-box modelling at the urban scale can be challenging. Black-box models can overcome these issues; however, they do not entail any information on buildings' physical characteristics, which is problematic especially for investigating retrofit scenarios or buildings' flexibility for demand response programs. To address the limitations of black-box as well as white-box models, hybrid modelling approaches using simplified and reduced order models are typically investigated to represent building characteristics with less inputs [16]. The lumped parameter thermal network is the most commonly used simplified model in building thermal energy modelling. In this model, the building components are modelled as an electrical circuit equivalent. The construction materials for each building components are modelled using thermal resistances (R-value) and thermal capacitances (C), if it has a thermal mass. The lumped parameters thermal network models each element in the building components as temperature-uniform element [26].

In the first order models, the entire thermal mass of the building is lumped to a single capacitance and no distinction is made between the structural mass and the indoor air mass. Second order models considered this difference by including a second capacitance, while third order models include three different capacitances for the envelope, the internal walls, and the indoor air, respectively. The fourth order models extended the third order by including a separate capacitance for the floor, while fifth order models have an additional capacitance for the roof [27]. Previous studies showed that reduced order models can capture the thermal dynamics of buildings and perform well in terms of short-term energy prediction [27, 28].

2.2 Estimating Existing Buildings Thermal Properties

To estimate the parameters of hybrid or gray-box models, different methods can be used depending on the order of magnitude of such models and available data. Table 1 provides a summary of these methods, the type of data they require, and the model parameters they can be used to estimate.

Table 1: Overview of the main methods used to identify buildings' thermal characteristics.

Estimation method	Required Data	Estimated Parameters
The Energy Signature (Balance point)	Heating system duty cycle, and outdoor temperature.	Resistance- Heating Power averaged for entire building (RK)
Degree Days	Energy consumption, indoor temperature, and outdoor temperature	Energy loss rate (Ɛ), and Thermal capacity for the entire building (C)
Decay Curves	Indoor and outdoor temperatures	The Thermal Time Constant for the whole building (RC)
Energy Balance	Heating system duty cycle, indoor and outdoor temperatures.	RC and RK

In the energy signature (balance point) method, the energy use of a single building is correlated to outdoor temperature. No dynamic behaviour can be captured with this method as it is assumed that the indoor temperature does not vary over the course of the day. Data from a single building is plotted on a graph, each point represents total energy load vs. mean outdoor temperature for a specific time period (daily is the most common). The reason this method is also called the balance point method is because most of such graphs show a distinct linear correction over a given temperature range, increasing with decreasing temperature during the heating season, and increasing with increasing temperature during the cooling season as shown in Figure 2. The temperature at which the heating load starts to rise is the heating balance point, and the temperature at which the cooling load starts to rise is the cooling balance point. In between these two temperatures is a temperature range over which energy use is relatively insensitive to outdoor temperature, because neither heating nor cooling is required to maintain a comfortable indoor temperature. The thermal resistance-heating power (RK) of a building can be approximated from the slopes of the lines on the graph, though the capacitance cannot be since no dynamic behaviour is captured within this method [29].

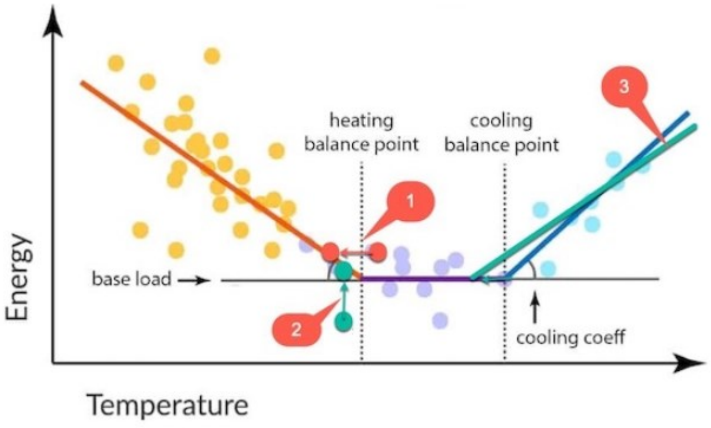


Figure 2: Typical Balance Point Graph [30].

The Degree Days method is commonly used to identify the heating and cooling energy requirements based on the integrated difference between base temperature and outdoor temperature. By using energy consumption data and the measured indoor temperature as the base temperature, the degree days method can quantify the correlation between energy consumption and the difference between the outdoor and indoor temperatures. By quantifying this correlation for time periods with heating inputs and an upward trend in the indoor temperature, the heating loss rate (Ɛ) and the whole building thermal capacity (C) can be estimated. The loss rate provides an indication of the insulation level and the building envelope tightness with a higher loss rate signifying a potential need for an envelope upgrade [31, 32].

The decay curve method can only be applied on free-floating periods (i.e. HVAC system is off) with the outdoor temperature is relatively constant, no solar gains, and minimal internal gains. For these specific periods, a thermal time constant can provide a proxy for how fast the building will achieve a new thermal equilibrium in response to changes in its internal and external thermal conditions (i.e. in the heating season, to cool from the prior setpoint to the setback setpoint). This method provides a measure that characterizes a building's ability to retain heat and represents the effects of the thermal inertia of building mass, which can be used to determine resistance-capacitance (RC) [29, 33].

To address the limitations of the decay curves method, the energy balance method can be implemented for outdoor air temperature variations and for the assumption that the building is not free-floating (i.e. building is in heating mode). However, filters must be applied to ensure that sufficient variations in the heating and indoor temperature are occurring at a given interval. For the filtered time periods, the Euler's method for numerical integration is used to solve the thermal energy balance equation for buildings and predict the indoor temperature over the specific period. The RC and RK parameters can then be estimated by minimizing the difference between the predicted indoor temperature values from the model and the measured temperatures using non-linear least-square optimization [29, 34]. Since the effect of internal gains and solar radiation as well as the lag in heating system response can significantly reduce the model's ability to properly estimate the RC and RK parameters, the numerical solution using Euler's method for numerical integration is used to drive the parameters.

2.2.1 Using temperature measurements in existing buildings to estimate model parameters

For many of the above-described methods, indoor and outdoor temperature measurements are needed to estimate the parameters of hybrid models. For example, Tabatabaei et al. and Van der Ham et al. used smart meter and thermostat data to estimate the heat loss rate and the thermal capacity of 99 and 67 residential buildings, respectively [31, 32]. They then used the heat loss rate to estimate thermal capacity and evaluate the thermal performance of residential buildings. The analysis showed that newer houses have lower heating loss rate, and the thermal capacity of the building envelope had a positive correlation with the size of the house. It is important to mention that this analysis used a small sample size and thus generalized conclusions could not be drawn.

John et al. used a dataset from smart thermostats to identify the thermal performance of approximately 10,000 buildings across North America using the decay curve method. They characterized seasonal variations of RC values and identified a significant correlation between RC values and building age [33]. Baasch et al. compared the energy balance and the decay curve method to estimate RC values from smart thermostats datasets by applying these methods on 2000 residential buildings in Canada and the US [29]. However, further validation of these methods and the relationship between RC values estimated using different methods for the same houses is needed across a larger sample of buildings. Furthermore, the effect of building attributes including their age, location and typology on estimated RC values was not investigated.

Chapter 3: Methodology

3.1 Data Description

The data used for this study was obtained from smart thermostat users who agreed to anonymously share their thermostat usage under the 'Donate Your Data (DYD)' program administered by ecobee Inc. [35]. The Dataset consists of five-minute interval data measured from the thermostat and sensors around the house, and users-reported metadata which describe the house characteristics. Outdoor weather data from the nearest weather station is also provided for each house at the same granularity as thermostat real-time data (i.e., 5-minute intervals). Table 2 provides details about DYD dataset.

Table 2: Ecobee Dataset Description

Attribute	Description
House ID	Anonymous unique ID of each user
HVAC mode	Indicates whether the HVAC system is off, heat, cool, auto, auxiliary heat
Indoor temperature (°F)	Measurement from ecobee thermostats
Outdoor temperature (°F)	Measurement from the nearest local weather station
Equipment runtime (seconds)	Measurement from ecobee thermostats
Country	User input (US, CA, etc.,)
Province/ State	User input (QC, ON, CA, FL, etc.,)
City	User input (Montreal, Toronto, New York, etc.,)
Building style	User input (Detached, Apartment, Row-House, others, etc,)
Floor area (ft^2)	User input (500, 1000, 1500, etc.,)
Number of Floors	User input (1, 2, etc.,)
Age of the house (years)	User input (10, 20, etc.,)

The data were collected from 95,215 houses across Canada and US between November 2015 and December 2019. The ASHRAE climate zone was identified for each house according to its geographical location. This study focused on investigating the thermal behaviour of residential buildings in colder climates, where the outdoor temperature is much lower than the indoor temperature and heating is required to maintain the indoor conditions at a comfort state. For that reason, the analysis focused on ASHRAE climate zones 4 to 8 (mixed, cool, cold, very cold, and subarctic\arctic) which include 60,003 houses. Figure 3 shows the available thermostat data for each ASHRAE climate zone. According to the figure, most of the houses belonged to climate zone 5 (57%) and climate zone 4 (27%). Climate zone 8 was not considered in this analysis since it only had five houses, which would not be representative.

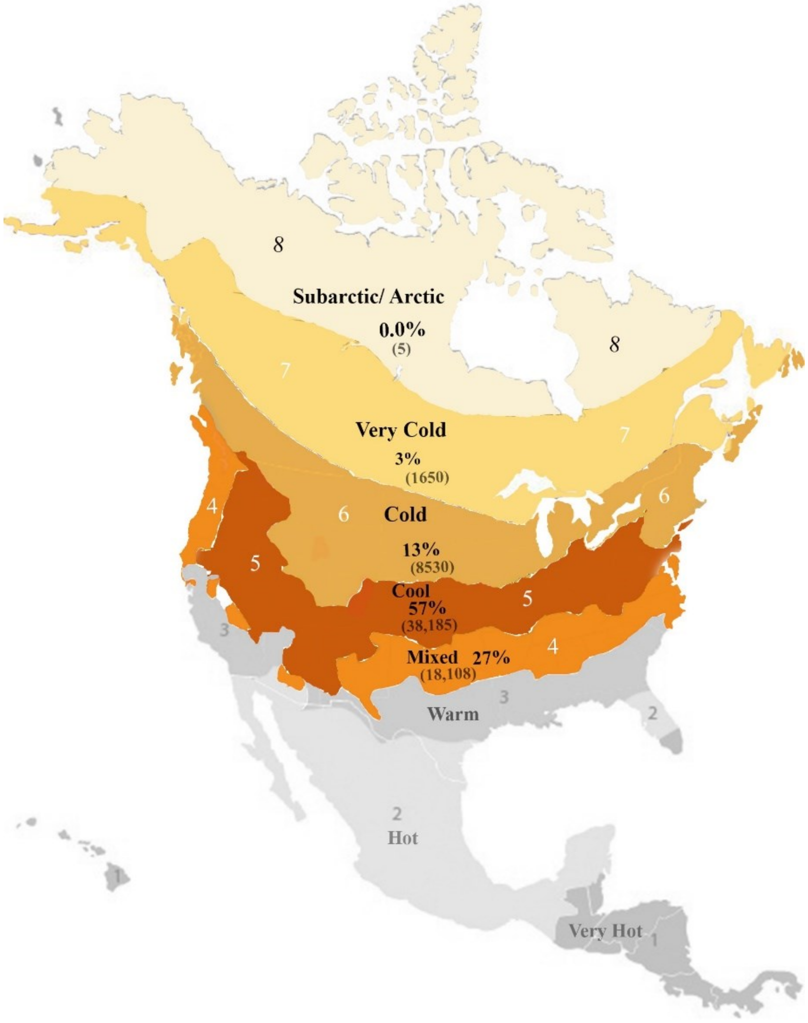


Figure 3: Insights on data availability in terms of ASHRAE climate zones.

Figure 4 shows the data availability with respect to building style. Around 53% of the data came from detached houses, while another 10.4% came from rowhouses, 6.4% from apartments and condos, and 2.2% from semi-detached houses. It is important to note that “building style”, “floor area”, and “building age” are user inputs, thus many users did not report this information or may have used inaccurate values which cannot be verified. In fact, approximately 26% of the houses did not contain information regarding “building style” as shown in Figure 4. Similarly, “floor area” and “building age” of each house were not provided for 12% and 21% of the houses, respectively. Approximately 2.5% of users entered very large floor areas (exceeding 5000 sq.ft), which we considered a user error (e.g., possibly representing the area of entire MURBs rather than individual units), thus they were excluded from the analysis.

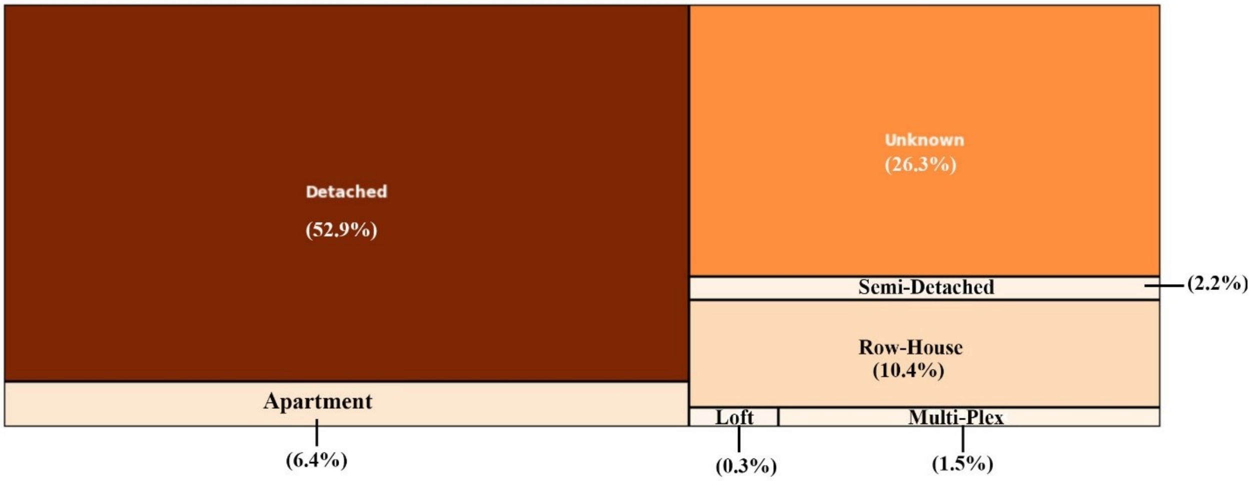


Figure 4: Insights on data availability in terms of building-styles.

3.2 Investigating Thermal Performance Using Grey-box Modelling Approaches

The main goal of this study is to investigate the thermal performance of residential buildings using the temperature readings from smart thermostat dataset. The time constant (RC value) is one the building’s thermal characteristics that can be estimated from the thermostat readings. This RC value is a measure that characterize the building ability to retain heat by quantifying the time the building takes to achieve a new thermal equilibrium after any change in its internal and/or external environmental conditions.

To estimate the buildings' time constants, the analysis focused only on periods at which the outdoor temperature was lower than the indoor temperature, no solar heat gains were expected, and other internal heat gain were minimal. These restrictions were necessary to ensure that fluctuation in indoor temperature were dominated by the indoor-outdoor temperature differential and/or the supplied heat from the heating system. Therefore, only night periods in colder months were considered, which were defined as periods between 8:00 pm to 5:00 am between October to April. These periods had a relatively larger difference between indoor and outdoor temperature, as well as a lower likelihood for additional disturbances such as opening windows in cold climates. Two different gray box modelling approaches were then used to estimate the time constant for each house:

i. The exponential decay curves of the indoor temperature.

ii. The numerical integration of the thermal energy balance equation.

The thermal energy balance in a building (expressed in equation 1) is the base of the physics equations used in both models.

$$C\frac{dT_{in}}{dt}(t) = \dot{Q}_{in}(t) + \dot{Q}_{h}(t) + \dot{Q}_{sol}(t) - \frac{1}{R}\left(T_{in}(t) - T_{ext}(t)\right) - \dot{Q}_{ven}(t) \tag{1}$$

Where:

- $\boldsymbol{T_{in}}$ is the indoor temperature measurement.

- $\boldsymbol{dt}$ is the time step.

- $\boldsymbol{T_{ext}}$ is the outdoor temperature.

- $\boldsymbol{\dot{Q}_{in}}$ is the internal heat gains.

- $\boldsymbol{\dot{Q}_{h}}$ is the heat flow supplied by heating system.

- $\boldsymbol{\dot{Q}_{sol}}$ is the solar radiation gains.

- $\boldsymbol{\dot{Q}_{ven}}$ is the heat flow due to ventilation.

- $\boldsymbol{C}$ is the lumped building capacitance.

- $\boldsymbol{R}$ is the lumped building thermal resistance.

However, since this analysis only consider night periods with no solar gains and with the assumption that the internal heat gain is minimum at night which makes the heat flow dominated by the heat supplied by the heating system and the heat flow due to the indoor and outdoor temperature difference, thus the thermal balance equation was rewritten as:

$$C\frac{dT_{in}}{dt}(t) = \dot{Q}_h(t) - \frac{1}{R}\left(T_{in}(t) - T_{ext}(t)\right) \qquad (2)$$

The heat flow supplied by heating system $\dot{Q}_h$ can be expressed as:

$$\dot{Q}_h(t) = \delta_{on}(t) \times K \qquad (3)$$

Where:

- $\boldsymbol{\delta_{on}}$ is the duty cycle for the heating system.

- $\boldsymbol{K}$ is the heating power.

The thermal energy balance therefore becomes:

$$\frac{dT_{in}}{dt}(t) = \frac{1}{RC}\left(T_{ext}(t) - T_{in}(t)\right) + \delta_{on}\, RK \tag{4}$$

Where:

- $\boldsymbol{RC}$ is the Resistance- Thermal capacitance for the entire building or the thermal time constant.
- $\boldsymbol{RK}$ is the Resistance- Heating Power averaged for entire building.

3.2.1 Least-squares Fitting of Decay Curves

The decay curve method assumes no heat input (i.e., HVAC is off) and that outdoor temperature is much lower than indoor temperature. At these times, the indoor temperature will decay towards the outdoor temperature at an exponential rate, which can be described by equation (5).

$$T_{in}(t) = (T_o - T_{ext})e^{\frac{-t}{RC}} + T_{ext} \tag{5}$$

Where:

- $\boldsymbol{T_o}$ is the initial indoor temperature.
- $\boldsymbol{T_{ext}}$ is the long-term final indoor temperature.
- $\boldsymbol{t}$ is the time elapsed in hours.

The RC parameter in this case represents the time it takes for a building's indoor temperature to realize 63.2% of the total change in its initial temperature given a constant outdoor temperature that is lower than the indoor temperature, while no (substantial) internal source of heat are active, as shown in Figure 5 [34].

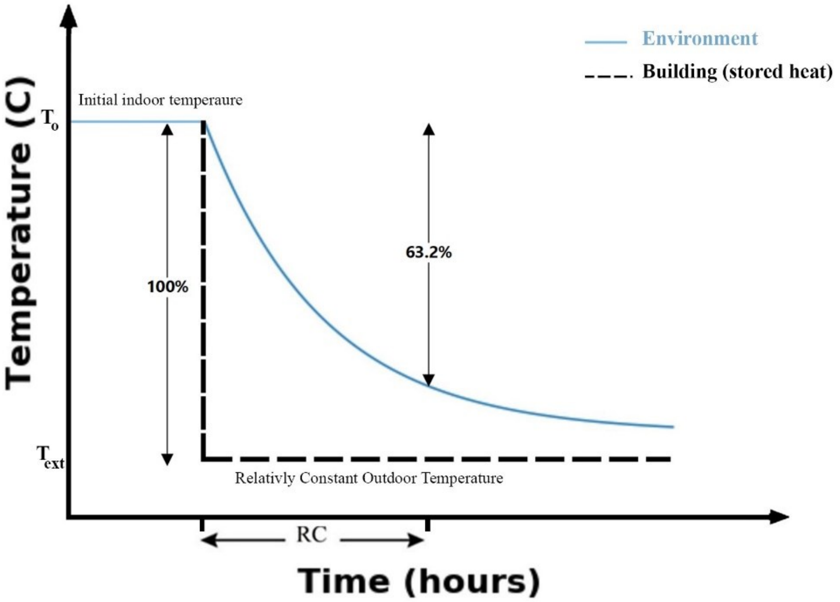


Figure 5: Conceptual representation of the time constant and decay curve adapted from [32].

Given the assumptions of the decay curve methods, time periods with free-floating conditions (i.e. when the HVAC is off and outdoor temperature is relatively constant) were first identified for each house where applicable. The analysis was then restricted to periods in which the difference between indoor and outdoor temperature was at least 5 ºF. Other filtering methods were also applied, which are summarized in Table 3.

Table 3: Filters applied to identify decay curves.

Filter	**Value**
The HVAC mode is off.	
Minimum drop in the internal temperature.	1 ºF
Minimum indoor-outdoor temperature difference.	5 ºF
Maximum change in the Outside temperature.	1 ºF

Based on these filters, over 55% of the houses did not have any eligible periods for the decay curve analysis method, while 14% of houses only had one decay curve (which were excluded since a minimum of two decay curves were deemed necessary to validate the estimated RC values). Figure 6 shows the frequency distribution of the number of decay curves found for each house. After the decay curves were identified, the initial temperature T_o and the time constant RC were estimated based on equation 1 using a non-linear least-square curve fitting approach. The average and standard deviation of RC values for each house was also calculated to remove unreliable values, which were deemed as those with a standard deviation higher than 25% of the mean.

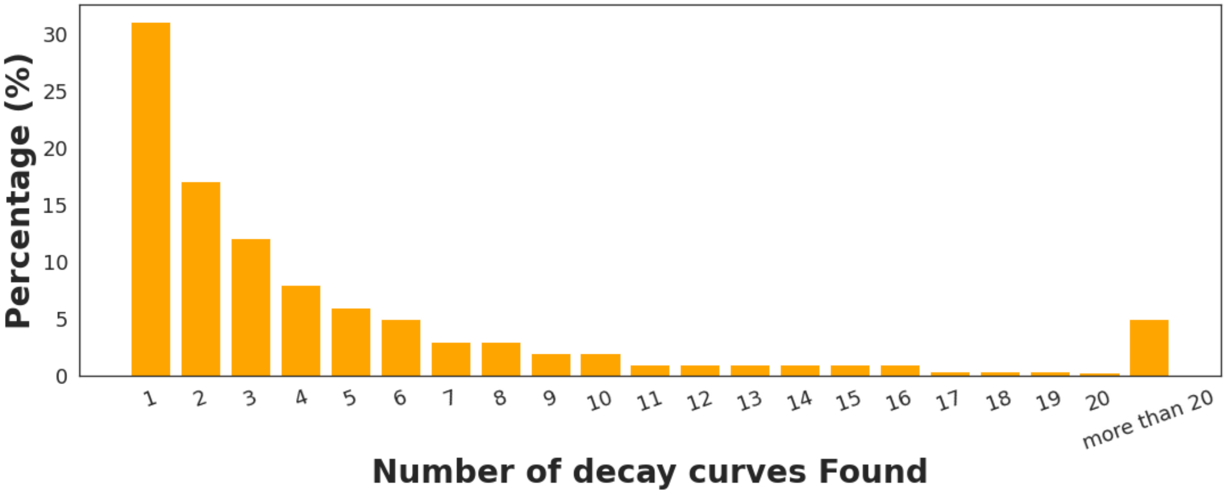


Figure 6: The distribution of identified decay curves per house.

3.2.2 Least-squares Fitting of the Numerically Integrated Thermal Energy Balance Equation

Another approach was investigated by solving a simplified energy balance equation (6) using Euler's method for numerical integration to estimate the parameters RC and RK as follows.

$$T_{in,i+1} = T_{in,i} + \Delta t \left[\frac{1}{RC}\left(\left(T_{ext,i} - T_{in,i}\right) + \delta_{on,i}\ RK\right)\right] \tag{6}$$

Where ***Δt*** is the timestep at which a change in the indoor temperature is detected.

This model was used to estimate RC and RK values to predict indoor temperature over specific time periods. Non-linear least-square optimization was then used to identify RC and RK values that minimize the differences between predicted values and measured indoor temperatures. This method was implemented to overcome the limitations of the decay curves approach in which the analysis was restricted to periods of constant outdoor temperature and no heating input in the system. The filters used for this method intended to identify periods in which heating was on for at least 5% of the time, and indoor temperature changed while outdoor temperature was consistently lower than indoor temperature by at least 2 ^{o}F. Other filters are summarized in Table 4, while Figure 7 shows the frequency distribution of the identified number of periods per house. To fit the model, ten periods were randomly selected for each house and fitted into the model to estimate RC and RK values. This process was repeated ten times followed by calculating the average RC values and the standard deviation for each house, and finally the unreliable results were removed (also defined as those with a standard deviation higher than 25%). The average RC values derived from each method for the same houses were then tested for correlation and plotted against each other.

Table 4: Filters applied to identify periods for the energy balance analysis method.

Filter	**Value**
The HVAC mode is 'heating'	
Minimum variance in the internal temperature	0.2 ^{o}F
Minimum indoor-outdoor temperature difference.	1 ^{o}F
Heating Duty cycle minimum.	5%

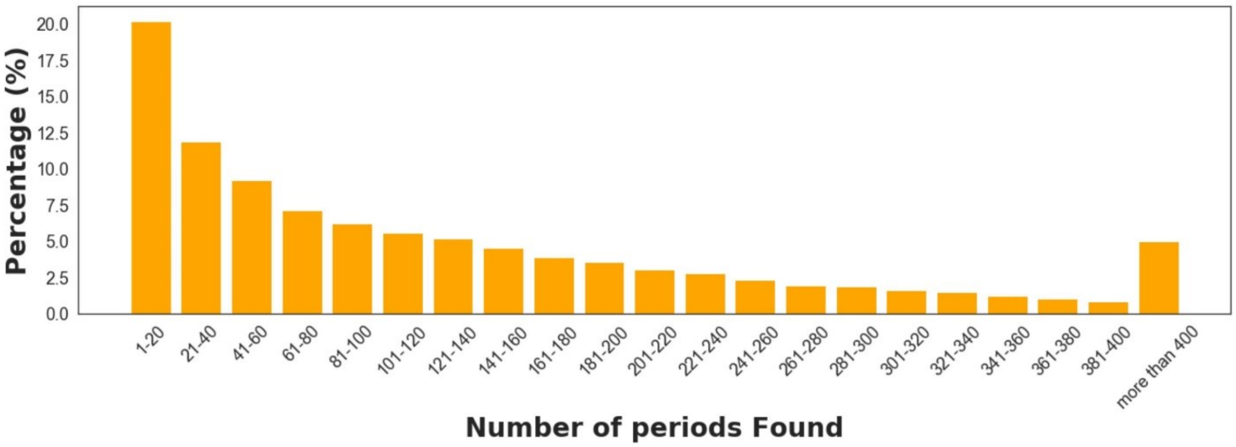


Figure 7: The distribution of identified periods per house for energy balance analysis.

3.2.3 Models Performance Evaluation

For the decay curve model, the R-squared was calculated for each identified period using equation (7). This R-squared values represent the goodness of fit for the predicted vs. actual indoor temperatures.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\widehat{T_{fit}} - T_{act})^2}{\sum_{i=1}^{n}(T_{act} - \bar{T})^2} \tag{7}$$

Where:

- $\boldsymbol{\hat{T}_{\iota}}$ is the fitted indoor temperature.
- $\boldsymbol{T_{act}}$ is the actual indoor temperature reading.
- $\boldsymbol{\bar{T}}$ is mean of the indoor temperature reading.

The average R-squared for the identified decay curves was 0.98, which indicates high accuracy in the fit of the curve to the actual indoor temperature readings. Figure 8 shows an example of the fitted curve versus the actual temperature measurements.

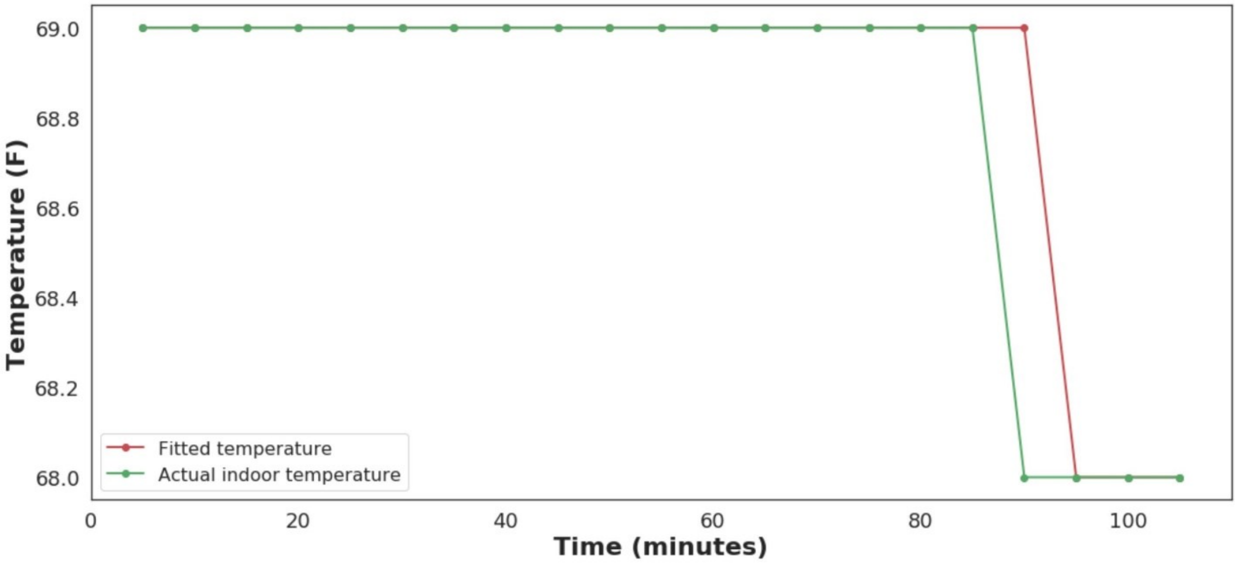


Figure 8: Fitting the indoor temperature for one decay curve.

On the other hand, the performance of the energy balance model was evaluated by comparing the predicted and the actual indoor temperatures for ten periods randomly selected and calculate the sum of squares of the residuals. The actual indoor temperature ranged from 285 to 295°K across the houses and the average sum of the squares was 6.3°K per period. Figure 9 shows the predicted values compared to the actual indoor temperature for one period as an example.

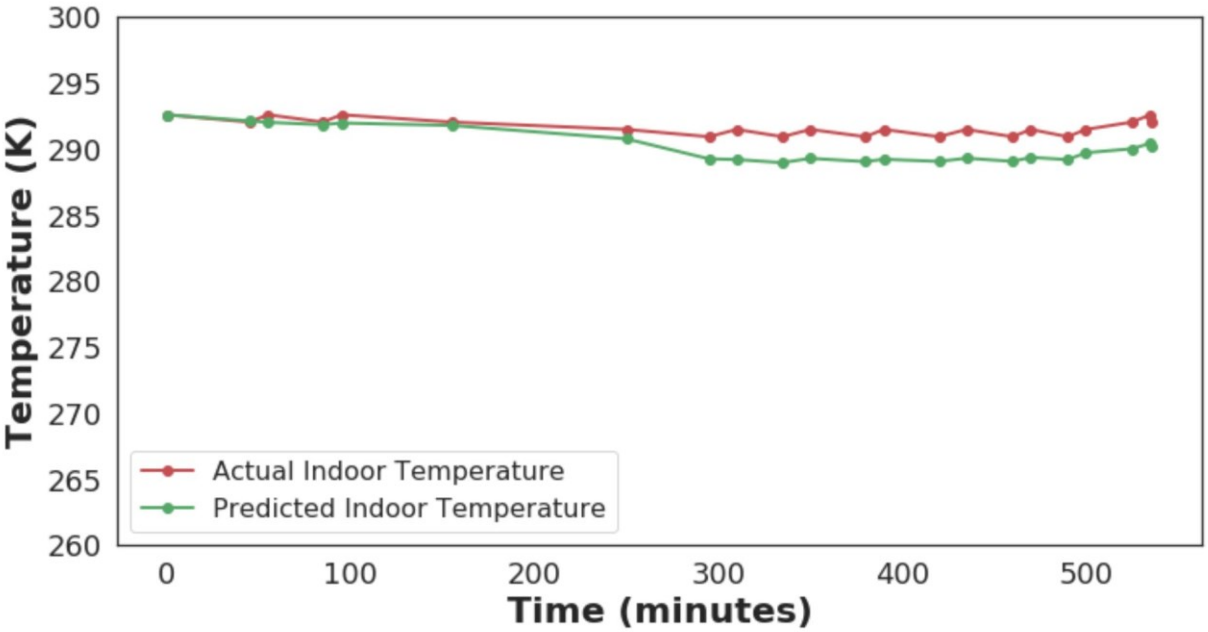


Figure 9: Comparing the predicted and the actual temperatures for one period from the energy balance model.

3.3 Estimating the Thermal Resistance (R-value) from the Estimated Time Constants

The obtained RC values were used to estimate corresponding R-values after making additional assumptions. First, RC-averaged values for each house were divided by a thermal capacitance C value of 10,000 and 20,000 Wh/K (which approximately represent the range of thermal mass of lightweight to heavyweight constructions [29], [36-41]). Results were then multiplied by the buildings' exposed surface areas, which were assumed to be 100, 200, 500, 1000 m^2 for apartments, row-houses, semi-detached and detached houses, respectively. Since the ecobee dataset did not include information on the thermal mass of the envelopes of each house nor the exposed surface areas, these assumptions were made based on existing literature references and typical dimensions of different house styles. Results of this analysis provided an approximate upper and lower bound of R-values (m^{2}K/W) for each house.

A benchmark for these R-values was also calculated based on ASHRAE 90.2 requirements which focus on the energy-efficient design of low-rise residential buildings. Assuming 14% window to wall ratio which is prescribed in the ASHRAE standard, a roof surface area covering 25% of total exposed area, and walls' surface area representing the remaining 75%, the effective envelope R-values based on the code requirements were found to range between 13.1 – 15.3 m^2K/W for climate zones 4 – 8. Assuming additional heat losses due to infiltration at a rate of 0.25 $L/s.m^2$, the effective code-compliant R-values would range between 7.7 – 9.0 m^2K/W for climate zones 4 – 8.

For further investigation, another benchmark was calculated for the Canadian houses based on part 9 of the Canadian National Building Code (NBC) which focus on low-rise residential buildings requirements. The effective envelope R-values based on code requirement were found to range between 12.2 – 14.7 m^2K/W for climate zones 4 – 8. Assuming additional heat losses due to infiltration at a rate of 0.25 $L/s.m^2$, the effective code-compliant R-values would range between 7.4 – 8.3 m^2K/W for climate zones 4 – 8. However, it must be noted that these estimated effective R-values represent the latest code requirements, thus actual effective R-values estimated using the ecobee dataset can be significantly lower due to poor craftsmanship, as well as in older houses which do not adhere to the latest ASHRAE and NBC requirements.

3.4 Investigating the Relationship between the Estimated Time-constants and Buildings' Attributes

The effect of building-age, floor-area, building-style, number of floors and ASHRAE climate zones on the identified RC values for each house was analyzed. First the houses were grouped with respect to the attributes reported in the metadata. For building age, each group represented a decade (ten years bins) while for the floor area, bins of 1000 ft2 were used. Additionally, buildings were grouped according to their style (apartment, row-house, semi-detached, and detached) and ASHRAE climate zones 4-7 (mixed, cool, cold and very cold). For this analysis, groups with less than 50 houses were not considered. Since the identified RC values for different houses did not follow a normal distribution according to Sharpio-Wilk's normality test ($p < 0.05$), non-parametric statistical tests were used throughout this analysis. The Kruskal-Wallis H test was used to investigate if there is any significant difference among the groups. If statistically significant differences were shown, pairwise comparisons using the Dunn procedure with a Bonferroni correction was applied [42].

3.5 Developing a Multi-class Classification Model to Predict the RC Range for Residential Buildings Using the Buildings' Attributes

The estimated RC values from the decay curve and the energy balance methods were grouped into categories (classes) using fifteen-hour bins. As a result, eleven classes were generated for all the houses with the higher class being more than 150 hours. Different approaches were then used in order to develop a classification model to predict the RC range for each house based on its attributes. The proposed methodology to develop the multi-class classification model is illustrated in Figure 10. The building's attributes used in this section are the building-style, building age, floor-area, ASHRAE climate zone, and number of floors. This methodology was applied to the RC values obtained from the decay curve and the energy balance methods separately. In the next subsections, each step of the workflow is explained in detail.

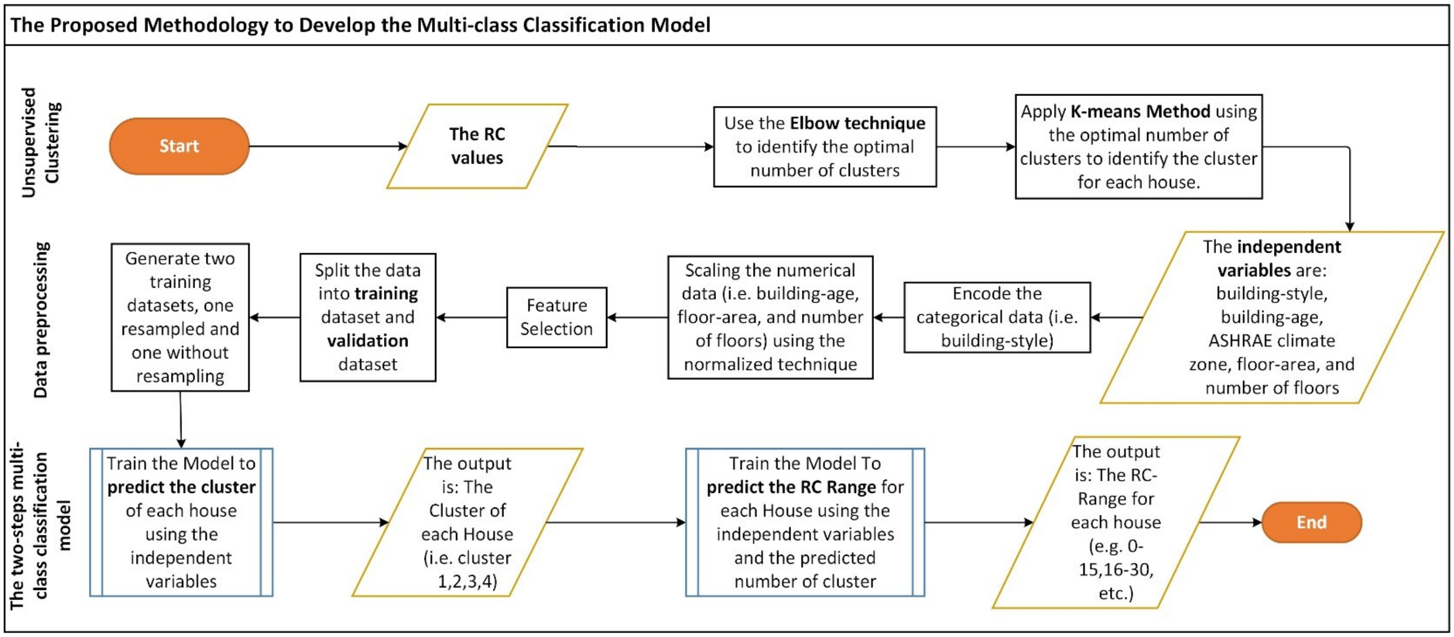

The Proposed Methodology to Develop the Multi-class Classification Model
Unsupervised Clustering
Start
The RC values
Use the Elbow technique to identify the optimal number of clusters
Apply K-means Method using the optimal number of clusters to identify the cluster for each house.
Data preprocessing
The independent variables are: building-style, building-age, ASHRAE climate zone, floor-area, and number of floors
Encode the categorical data (i.e. building-style)
Scaling the numerical data (i.e. building-age, floor-area, and number of floors) using the normalized technique
Feature Selection
Split the data into training dataset and validation dataset
Generate two training datasets, one resampled and one without resampling
The two-steps multi-class classification model
Train the Model to predict the cluster of each house using the independent variables
The output is: The Cluster of each House (i.e. cluster 1,2,3,4)
Train the Model To predict the RC Range for each House using the independent variables and the predicted number of cluster
The output is: The RC-Range for each house (e.g. 0-15,16-30, etc.)
End

Figure 10: The workflow of the developing the multi-class classification model of RC values based on building attributes

3.5.1 Unsupervised Clustering of the Estimated Time-constants

The RC values estimated from the decay curve and the energy balance methods were clustered using the K-means method. This type of clustering is called the unsupervised clustering since no label was given to the clustering algorithm in order to use it to define the structure of clusters. The optimal number of clusters was found by using the Elbow method with respect to the within cluster sum of squares (WCSS). The WCSS can be calculated using equation 8.

$$WCSS = \sum_{C_k}^{C_n} \sum_{d_i in\, C_i}^{d_m} (distance(d_i\,, Ct_k)^2 \tag{8}$$

Where:

- **Ct** is cluster centroid.

- **d** is the data point in each cluster.

The WCSS will keep decreasing with the increase of the number of clusters till it reached zero when the number of clusters is equal to the number of datapoints in the dataset. The goal of the elbow method is to find the lower number of clusters that minimize the WCSS. Figure 11 shows the elbow method using the RC values from the decay curve and the energy balance methods. As shown in the figure, the optimal number of clusters, which can be seen as the inflection point in the curve, is four clusters for both the decay curve and the energy balance RC values. After finding the optimal number of clusters, the K-means method was applied to find the right cluster for each house. Table 5 shows the description of the clusters in terms of the RC classes in each cluster.

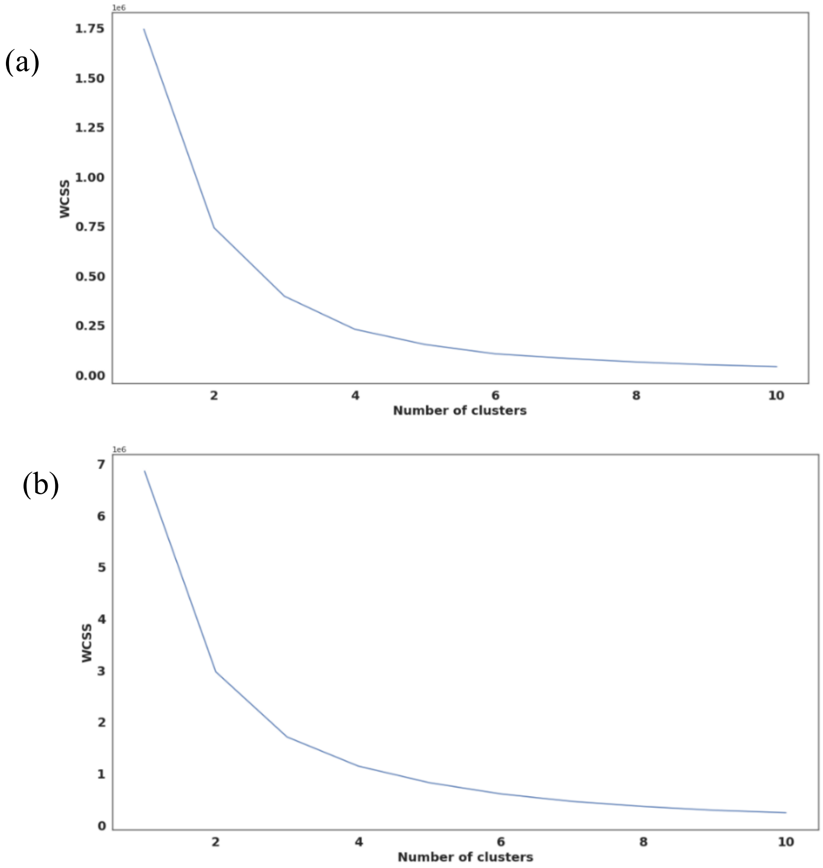


Figure 11: The Elbow method for: (a) the decay curve RC values, and (b) the energy balance RC values.

Table 5: Description of the clusters.

Decay Curve		
Cluster	**Number of classes in each cluster**	**RC Classes in each cluster**
1	2	0-15 and 16-30
2	2	31-45 and 46-60
3	3	61-75, 76-90, and 91-105
4	4	106-120, 121-135, 136-150, and more than 150
Energy Balance		
Cluster	**Number of classes in each cluster**	**RC Classes in each cluster**
1	3	0-15, 16-30, and 31-45
2	1	46-60
3	2	61-75 and 76-90
4	5	91-105, 106-120, 121-135, 136-150, and more than 150

3.5.2 Data Preprocessing

3.5.2.1 Encode the Categorical Data

All Machine learning algorithms (ML) requires numerical data, thus, the first step in data pre-processing was to encode the categorical features (i.e. building styles) into numerical values using the one hot encoding technique. One hot encoding transforms a single variable with n observations and d distinct values, to d binary variables with n observations each. Each observation indicating the presence (1) or absence (0) of the dichotomous binary variable.

3.5.2.2 Feature Scaling

The second step in data pre-processing was feature scaling of the numerical features (i.e. building age, floor area, and number of floors) using the normalization technique. The numerical features used to describe the building's attributes have a different dynamic range. Thus, feature normalization was required to approximately equalize ranges of the features such that they have approximately the same effect in the predictions process.

3.5.2.3 Feature Selection

To select the optimal number of features (i.e. the independent variables) for the classification model, the importance of the building-style, building-age, floor-area, ASHRAE climate zone, and number of floors in predicting the range of the RC value for each house was ranked using the importance factors from the Adaptive boosting algorithm (AdaBoost). The AdaBoost algorithm was proposed by Yoav Freund and Robert Shapire in 1995 for generating a strong classifier from a set of weak classifiers (i.e. base learners) [43]. For this study, a total of fifty random forest classifiers (RF) were used to predict the RC range for each house. The AdaBoost algorithm started by assigning equal weights to each house in the dataset. These weights can be calculated using equation 9.

$$W_i = \frac{1}{N} \tag{9}$$

Where:

- $\boldsymbol{W_i}$ is the initial weight given to each house in the dataset.
- **N** is the number of the houses in the dataset.

The RF classifiers were then fitted on the dataset sequentially. Based on the performance of each classifier the weights assigned to the houses were adjusted by decreasing the weights for the correctly classified houses and increasing the weights for the incorrectly classified houses. With this weight's adjustment, the subsequent classifier can focus more on the difficult data sample (i.e. the houses with higher weights). After fitting all the RF classifiers, the importance factors of each building attributes were examined. The importance feature is computed as the normalized total reduction of the criterion brought by that attribute. The higher the factor is the more influence the attribute has on predicting the RC range. Table 5 shows the ranking of the building attributes based on the importance factors for the decay curve and the energy balance results. As seen in the table, the importance factors of the building's attributes were relatively close in both methods. As a result, the five attributes were used as the independent variables to predict the RC ranges. For the RC values obtained using the decay curve method, floor area and ASHRAE climate zone was found to be the most influential attributes on predicting the RC ranges. While for RC values from the energy balance methods, ASHRAE climate zone and building age were the most influential attributes.

Table 6: Ranking of Building Attributes based on importance.

Rank	Building's Attributes	Importance Factor
	Decay Curve	
1	Floor area	0.296
2	ASHRAE climate zone	0.287
3	Number of floors	0.256
4	Building age	0.233
5	Building Style	0.221
	Energy Balance	
1	ASHRAE climate zone	0.161
2	Building age	0.143
3	Number of floors	0.139
4	Floor area	0.134
5	Building Style	0.124

3.5.2.4 Generating the Training and Validation Datasets

The last step in data pre-processing was splitting the dataset into training and validation subsets. The records for each subset were chosen randomly from the entire data set. The training subset was populated using 75% of the data, while the remaining 25% were used for validation.

The training subset included all the eleven RC classes, however, the number of houses in some of the classes (i.e. the minority classes) was much lower than other classes (i.e. the majority classes), which might cause imbalance training subset. To avoid this problem, another training subset were generated by applying resampling techniques on the original training datasets. Two resampling techniques were applied in this study namely, 1) the under-sampling technique and 2) the synthetic minority oversampling technique (SMOTE). In the SMOTE technique, the minority classes were oversampled by finding the K-nearest-neighbours (KNN) for randomly chosen observations (i.e. houses) to generate similar ones, while randomly chosen observations were deleted from the majority classes in the under-sampling technique. The distribution of the training datasets before and after resampling can be found in Appendix 14.

It is important to mention that the four training subsets from before and after resampling of the decay curve and energy balance results were used to train the multi-class classification model separately.

3.5.3 The Two-step Multi-class Classification Model

The prediction of the RC ranges was made on two steps, first by training the model to predict the cluster of the house using the building's attributes. After this step, the cluster number and the building's attributes were used to predict the RC range. These steps are illustrated in Figure 12. The ML algorithm used for the two-step classification models was the AdaBoost algorithm with RF as base learner. A total of fifty RF models were used, each consisting of fifty decision trees. This structure was found to be the optimum structure after multiple trials. Using more RF models or decision trees caused the model to be overfitted, while using fewer models decreased the accuracy of the model predictions.

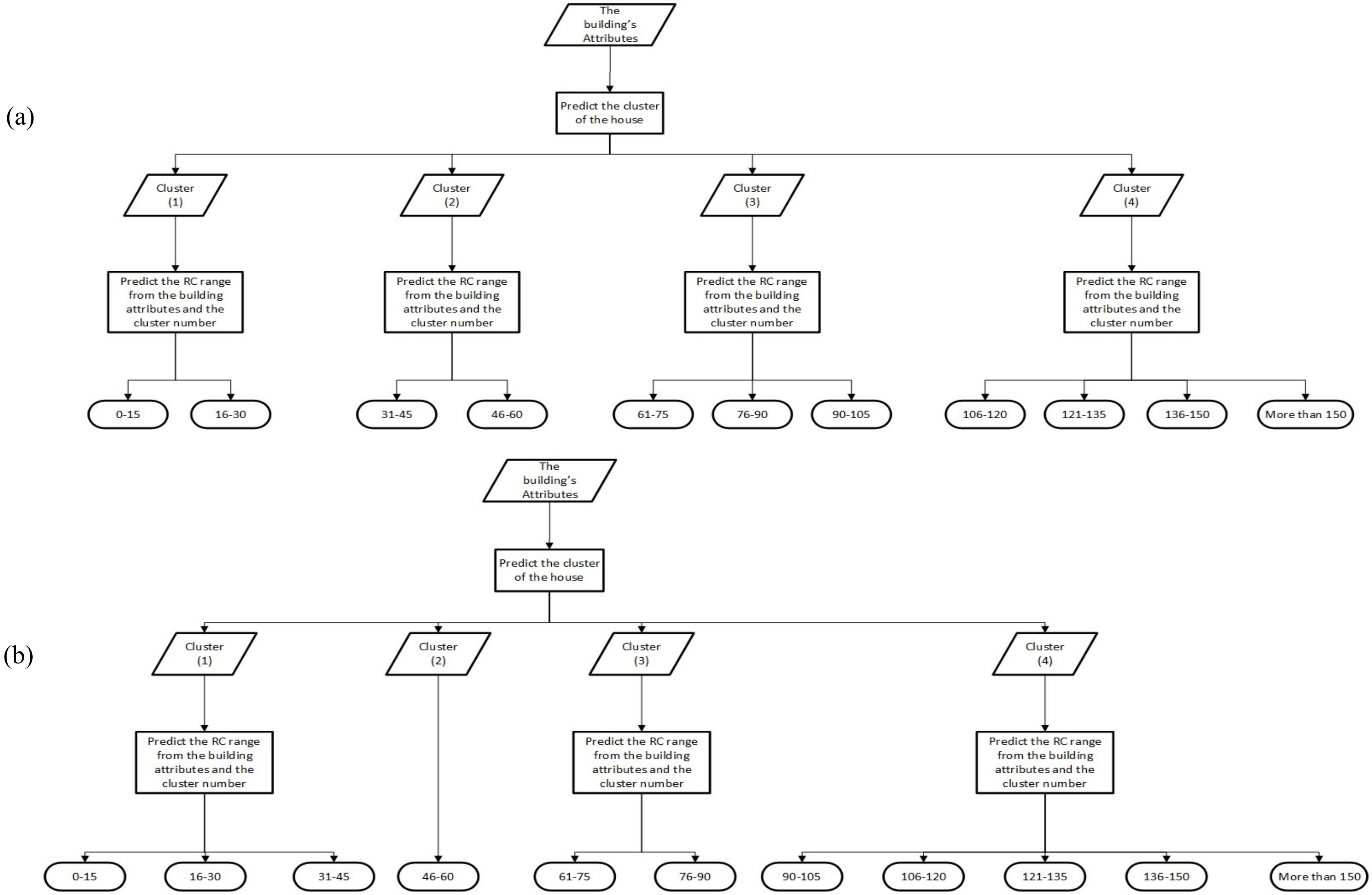

(a)
The building's Attributes
Predict the cluster of the house
Cluster (1)
Cluster (2)
Cluster (3)
Cluster (4)
Predict the RC range from the building attributes and the cluster number
Predict the RC range from the building attributes and the cluster number
Predict the RC range from the building attributes and the cluster number
Predict the RC range from the building attributes and the cluster number
0-15
16-30
31-45
46-60
61-75
76-90
90-105
106-120
121-135
136-150
More than 150
(b)
The building's Attributes
Predict the cluster of the house
Cluster (1)
Cluster (2)
Cluster (3)
Cluster (4)
Predict the RC range from the building attributes and the cluster number
Predict the RC range from the building attributes and the cluster number
Predict the RC range from the building attributes and the cluster number
0-15
16-30
31-45
46-60
61-75
76-90
90-105
106-120
121-135
136-150
More than 150

Figure 12: The workflow of the two-step multi-class classification model for: (a) the decay curve results, and (b) the energy balance results

3.5.4 Evaluating the performance of the Multi-class Classification Model

The accuracy of the two-step multi-class classification model was evaluated by making predictions against the validation subset. Accuracy was measured by comparing the predicted RC range from the model to the actual RC range of each house in the validation subset, then providing a confusion matrix. The confusion matrix was then analysed to estimate the accuracy and misclassification rate. Moreover, the relationship between the actual values of RC range in the validation subset and the predicted values was assessed using the Kolmogorov–Smirnov test (KS test). The KS test reported the maximum difference between the distributions of the actual and predicted values. Based on the KS results, the influence of resampling the training subsets on the accuracy of the model predictions was examined.

Chapter 4: Results and Analysis

4.1 The Time Constant (RC values) Results

For the decay curves method, 2112 houses with reliable results were found. The median of the identified RC values for these houses was 52 hours, and the mean was 58 hours, while 90% of the values ranged between 11– 80 hours. On the other hand, the energy balance model had 21,921 houses with reliable results. The median of identified RC values for these houses was 51 hours, and the mean was 55 hours, while 90% of the values ranged between 11 – 110 hours. Figure 13 shows the frequency distribution of the identified RC values from both models, while a full descriptive statistic of the RC values from both methods with respect to the building's attributes can be seen in appendix 1- 4.

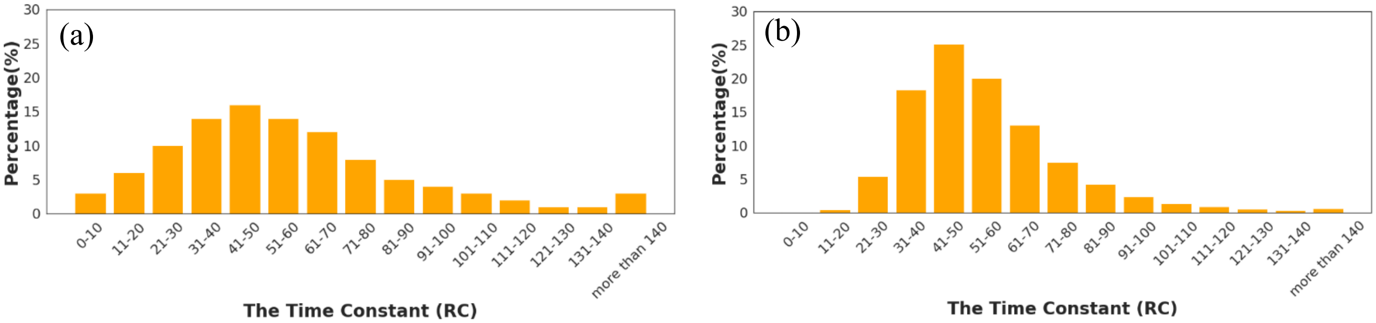


Figure 13: The distribution of RC values identified using: (a) the decay curves model, (b) the energy balance model.

4.2 The R-value Results

To gain a better understanding of these estimated RC values, equivalent minimum and maximum R-values were calculated for each house based on the methodology described in section 3.3. These estimated minimum R-values had a median of 2 m^{2}K/W for the decay curve and the energy balance results, with more than 85% of the values ranging between 0.5 to 4 m^{2}K/W.

On the other hand, the median for estimated maximum R-values was 4 m^{2}K/W for the decay curve and the energy balance results. More than 85% of the values form both methods ranged between 1 to 14 m^{2}K/W. The distributions of both minimum and maximum R-values are shown in Figure 14, while the distribution with respect to the building's attributes can be seen in appendix 5- 8.

The benchmark range of code-compliant R-values calculated in the previous section was 13.1 -15.4 m^{2}K/W and 7.7 -9.0 m^{2}K/W after accounting for infiltration. Although the estimated R-values were relatively lower for the majority of the houses than the estimated code benchmark, this could be due to poor workmanship in actual buildings. Furthermore, many houses in the ecobee dataset were likely built to older (and less stringent) code requirements and their envelopes likely deteriorated over time, which resulted in a lower thermal performance.

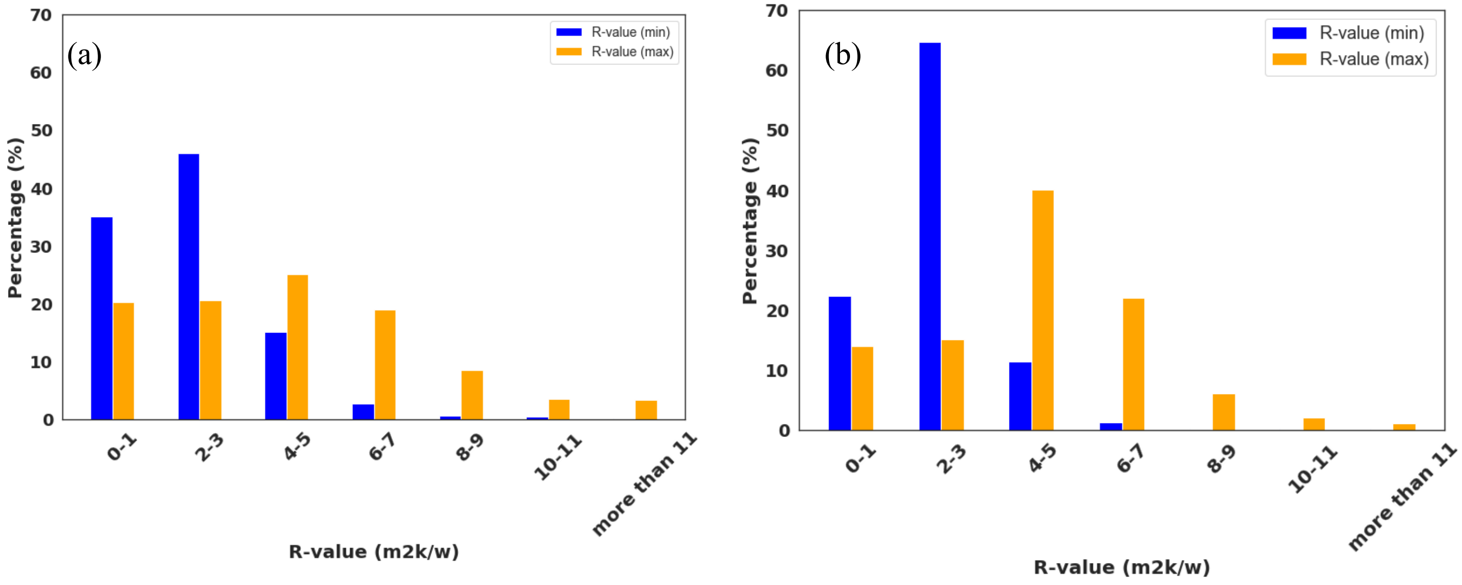


Figure 14: The distribution of the estimated minimum and maximum R-values based on: (a) the decay curve model, (b) the energy balance model

4.2.1 R-value Results for Canadian Houses

The estimated minimum R-values for the Canadian houses had a median of 2 m^{2}K/W for the decay curve and 3 m^{2}K/W the energy balance results, with more than 80% of the values ranging between 0.5 to 4 m^{2}K/W.

Furthermore, the median for estimated maximum R-values was 5 m^{2}K/W for the decay curve and the energy balance results. More than 80% of the values form both methods ranged between 1 to 7 m^{2}K/W. The distributions of both minimum and maximum R-values are shown in Figure 15. Comparing to the NBC-code-compliant R-values which ranged from 7.4 to 8.3 m^{2}K/W after accounting the inflation, the estimated R-values for the majority of the Canadian houses were lower than the code

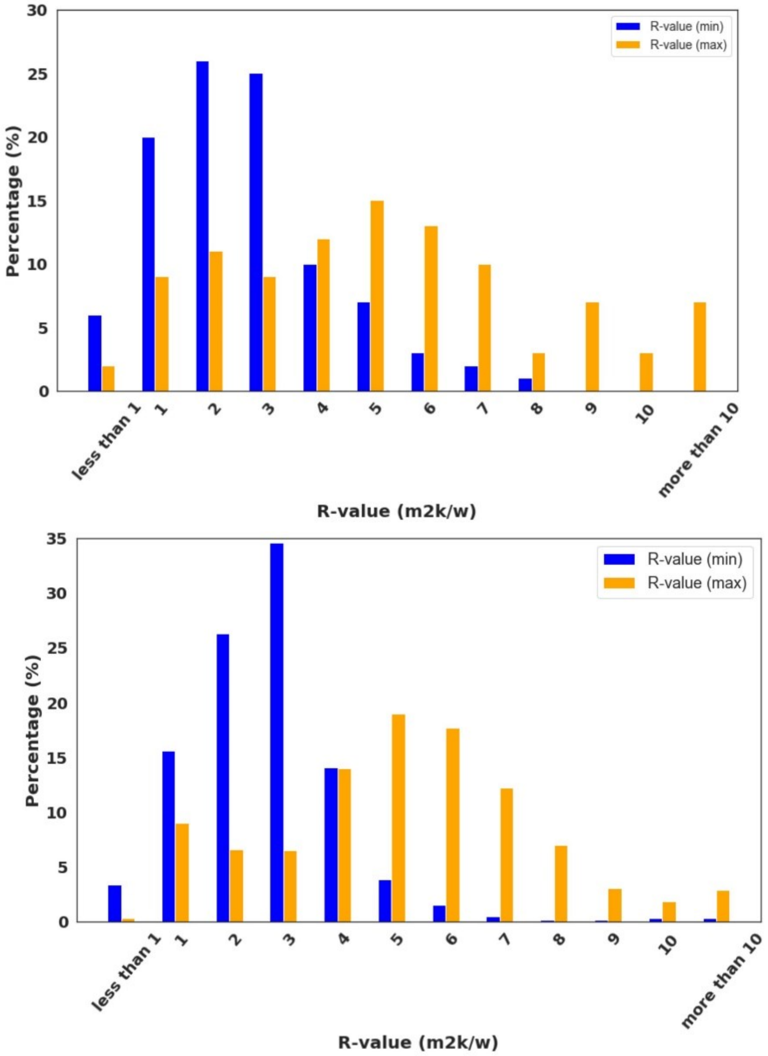


Figure 15: The distribution of the estimated minimum and maximum R-values for the Canadian houses based on: (a) the decay curve model, (b) the energy balance model

4.3 Models Comparison

Results from both the decay curves and energy balance methods were compared against each other. The number of houses with eligible periods for each method, as well as the percentage of excluded (unreliable) houses are compared in Table 7.

Table 7: Comparison between the decay curves and energy balance methods.

Method	Result	
Decay curve	Total number of houses identified	26,862
	Percentage of unreliable results	92.14%
	Median RC value	52 hours
Energy balance	Total number of houses identified	52,683
	Percentage of unreliable results	58.39%
	Median RC value	51 hours

Further analysis of the differences between both methods entailed comparing RC values for the subset of houses in which estimates were made using both methods, as shown in Figure 16. This analysis only focused on 1014 houses, (which were those for which suitable periods were identified to estimate RC values using both methods). A statistically significant positive correlation was found between the RC values obtained using both methods (r2 = 0.4, $p<0.05$). This observation supports the reliability of results obtained using both models but highlights some of their discrepancies given that both only relied on statistical inference only to estimate a proxy for thermal performance.

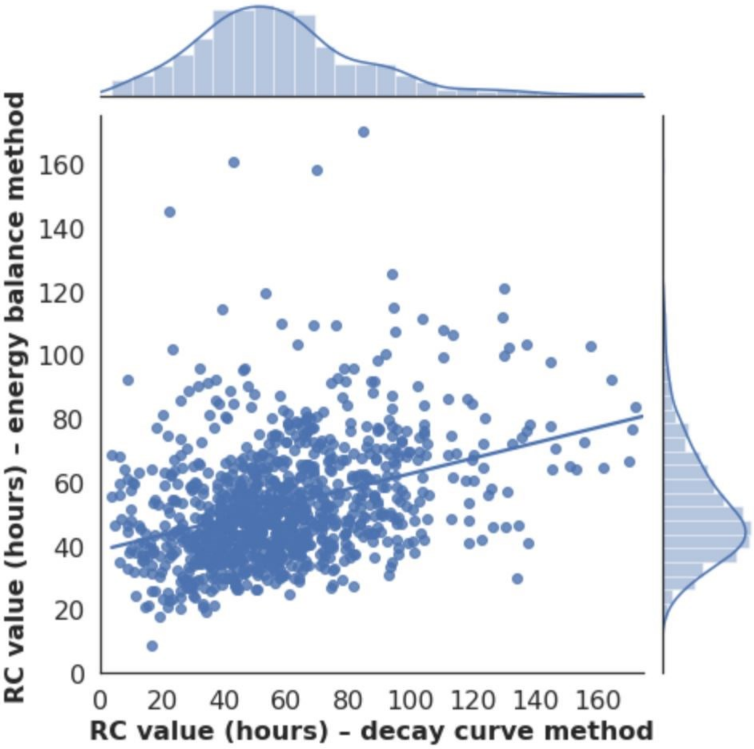


Figure 16: Linear correlation between RC values calculated using the decay curves and the energy balance models.

4.4 Investigating the Relationship between RC-values and Building Attributes

The Kruskal-Wallis H test showed a significant difference in median RC values obtained from both methods among the different building-ages, building-styles, and the ASHRAE climate zone groups as shown in Figure 17. The results of the Kruskal-Wallis H test and the Dunn pair-comparison can be seen in appendix 9- 13.

Figure 17a indicates that newer buildings (age <31) had the highest median RC values estimated from the energy balance model as they tend to have a better insulation level, while the middle-aged buildings (31-60) had the lowest median RC value. The decay curve RC values showed a similar result with the newer buildings (age<11) having the highest RC median and the older houses showing a decreasing trend. However, it is important to note some of the inconsistencies observed between RC values obtained using each method with respect to building age. Since a different set of houses was identified as applicable for each method, the sub-samples for each of decade included different houses for each method (e.g., for the energy balance method, 10,460 applicable houses were built 11-30 years ago compared to only 921 for the decay curve method, which may have skewed their median estimated RC values). Nevertheless, general observations regarding houses' thermal performance relative to building attributes were identified through this analysis. It is also worth mentioning that the results from some studies showed that older houses can have a slower change rate in the indoor temperature due to a better thermal mass than the newer houses [44], this observation aligned with the results from the energy balance method where the older houses have a high RC median. However, the results from the decay curve method can be due to material deterioration with aging.

Moreover, RC values obtained from both methods showed relatively similar trends with respect to other building attributes. For example, RC values generally increased in larger buildings as shown in Figure 17b, which was expected as a result of the greater thermal mass of these buildings. Another expected increasing trend could be seen based on climate zone using the energy balance and the decay curve results, confirming better thermal performance in colder climates as shown in Figure 17c. Finally, Figure 17d showed a decreasing trend in median RC values as the building's exposed area increased based on building style (i.e., highest RC values identified for apartments in the energy balance results and the rowhouses in the decay curve and lowest for detached houses). Since apartments typically have one surface area exposed to the outside and rowhouses typically have two, it is expected from these styles to retain heat longer than the detached houses with all surfaces exposed.

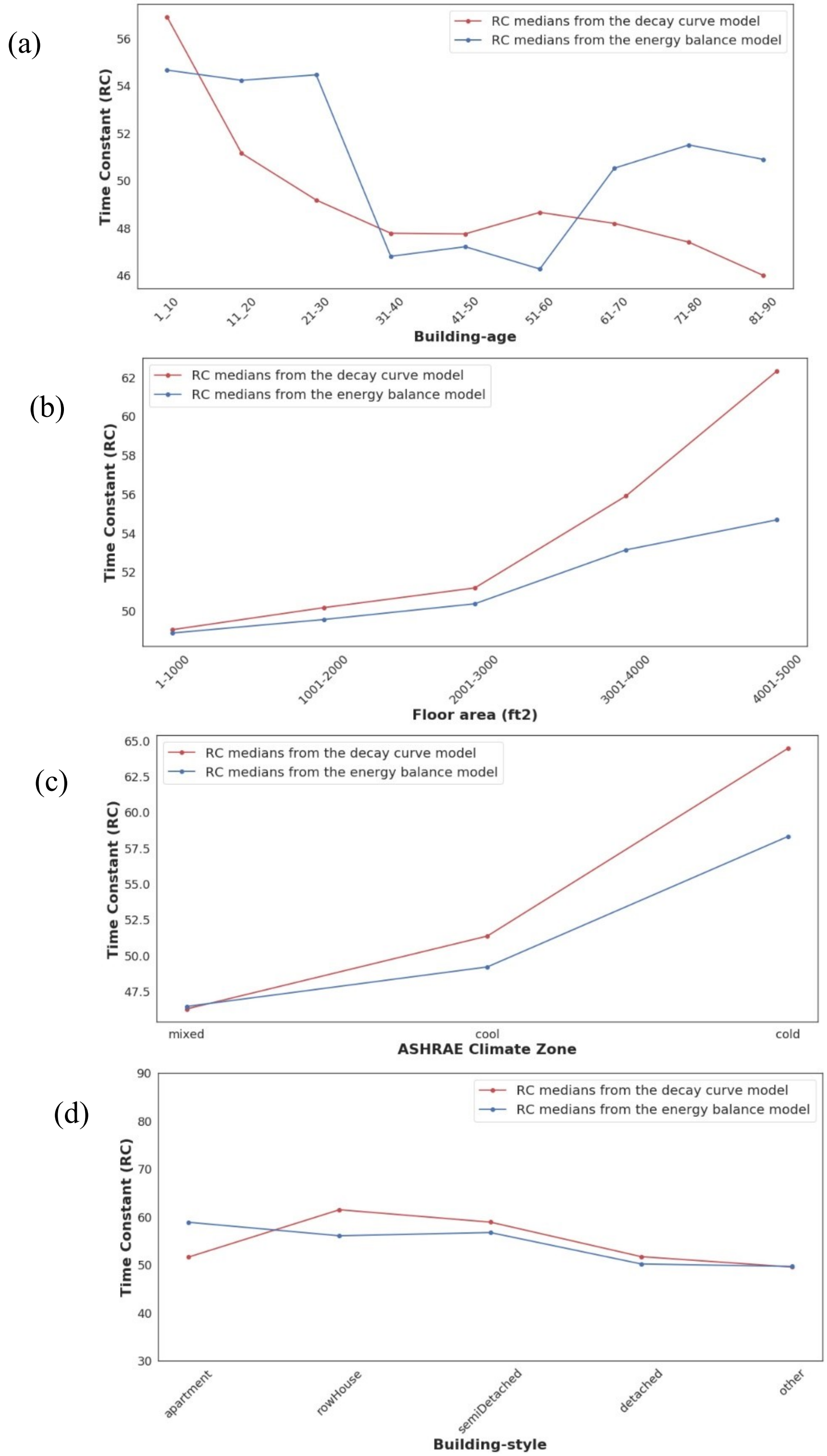


Figure 17:Estimated RC values based on the decay curve and energy balance methods for different: (a) building ages, (b) floor areas, (c) ASHRAE climate zones, and (d) building-styles.

4.4.1 Investigating the Distribution of the RC-values with respect to Building-style

The distribution of the building styles in the RC range categories is shown in Figure 18 for the decay curve and the energy balance RC values. From the figure, it can be seen that in the decay curve results the percentage of the detached houses was the highest when the RC values ranged between 81 and 90 hours. However, their percentage dropped when the RC values were more than 100 hours. Similar pattern was seen in the energy balance results, where the highest percentages of the detached houses were found in the RC categories between 31 to 60 hours followed by decreasing in the percentages for the categories with higher RC values. On the other hand, the percentage of the apartments was lower in the categories with lower RC values, and reached the highest level when the RC values were more than 130 and 120 hours in the decay curve and the energy balance results respectively. These observations validate the lower RC values for the houses with more exposed surface area to the outside (i.e. detached houses) and the higher RC for houses with less exposed surface area (i.e. apartments).

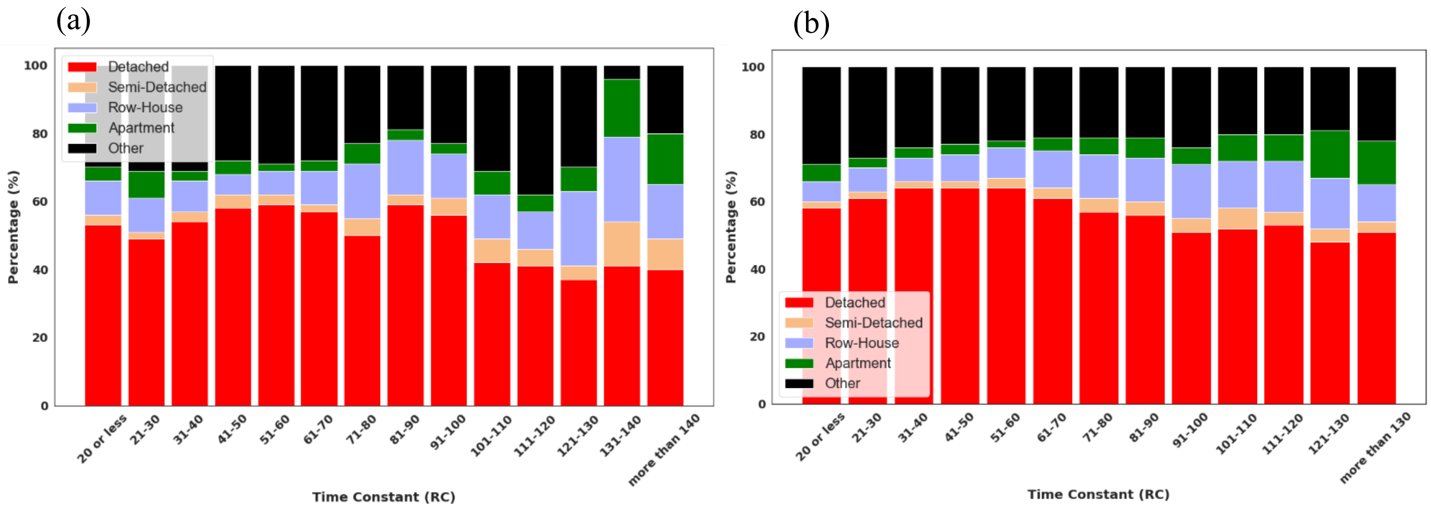


Figure 18: The distribution of each RC range category with respect to the building-style for: (a) the decay curve RC-values and (b) the energy balance RC-values.

4.4.2 Investigating the Distribution of the RC-values with respect to Building-age

Figure 19 shows the distribution of the building age groups in the RC ranges categories form the decay curve and the energy balance methods. For the energy balance results, an increasing trend can be seen in the percentages of the newer houses (age<11) reaching the highest percentage at the category with RC values ranged from 101 to 110 hours, which was also the highest category in the percentage of the newer houses in the decay curve results. This support the observation from the previous section, that newer houses tend to have higher RC values.

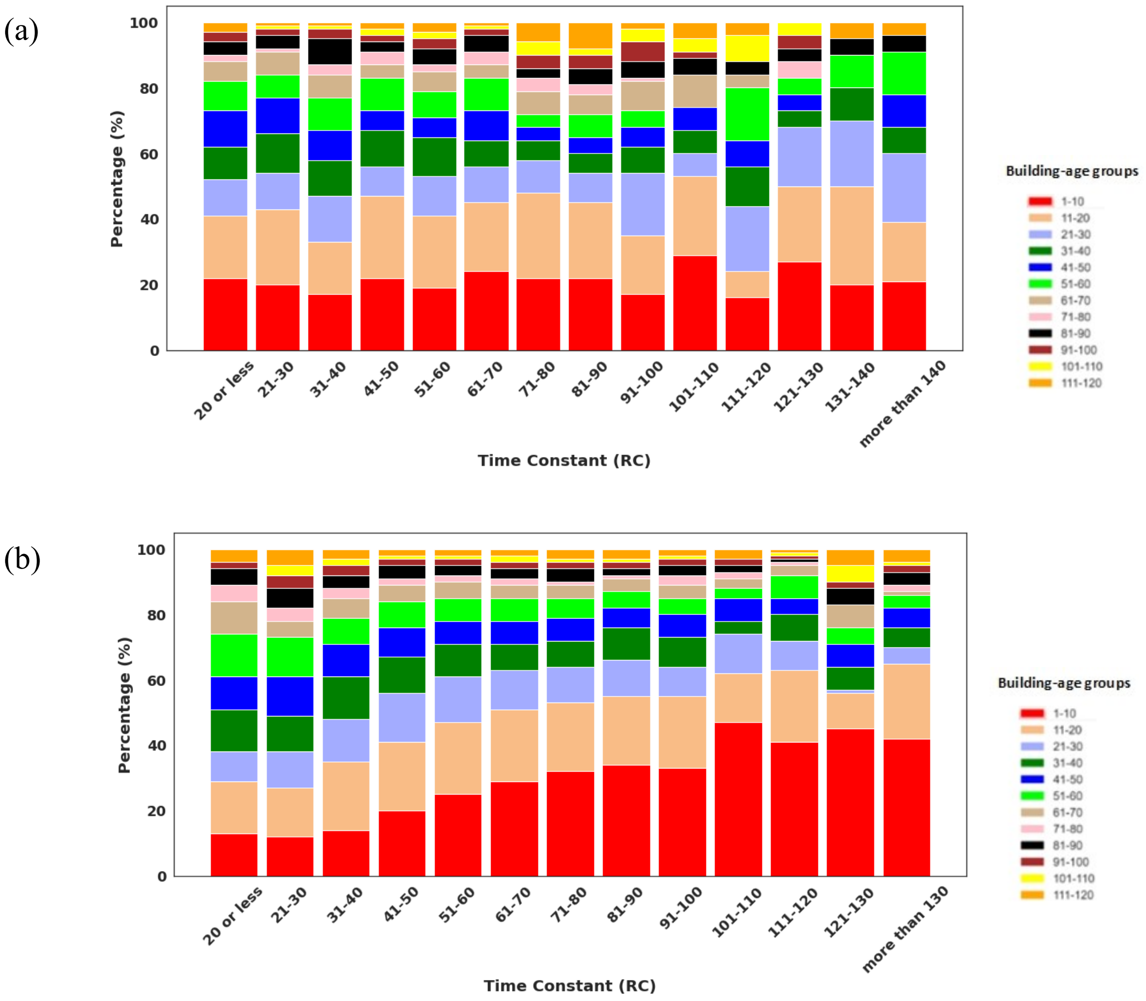


Figure 19: The distribution of each RC range category with respect to the building-age for: (a) the decay curve RC-values and (b) the energy balance RC-values.

4.4.3 Investigating the Distribution of the RC-values with respect to Floor-area

The distribution of the floor-area groups in the RC range categories from the decay curve and the energy balance results is shown in Figure 20. For smaller houses (area <1000 ft^2), the highest percentage was found in categories with RC values ranged from 131 to 140 hours for the decay curve results. However, for the energy balance results, the highest percentage of the smaller houses was in the category with 20 hours or less RC values. The larger houses (area >4000 ft^2) on the other hand had the highest percentage in the category with the RC values ranged from the 131 to 140 hours for the decay curve results and more than 130 hours for the energy balance results.

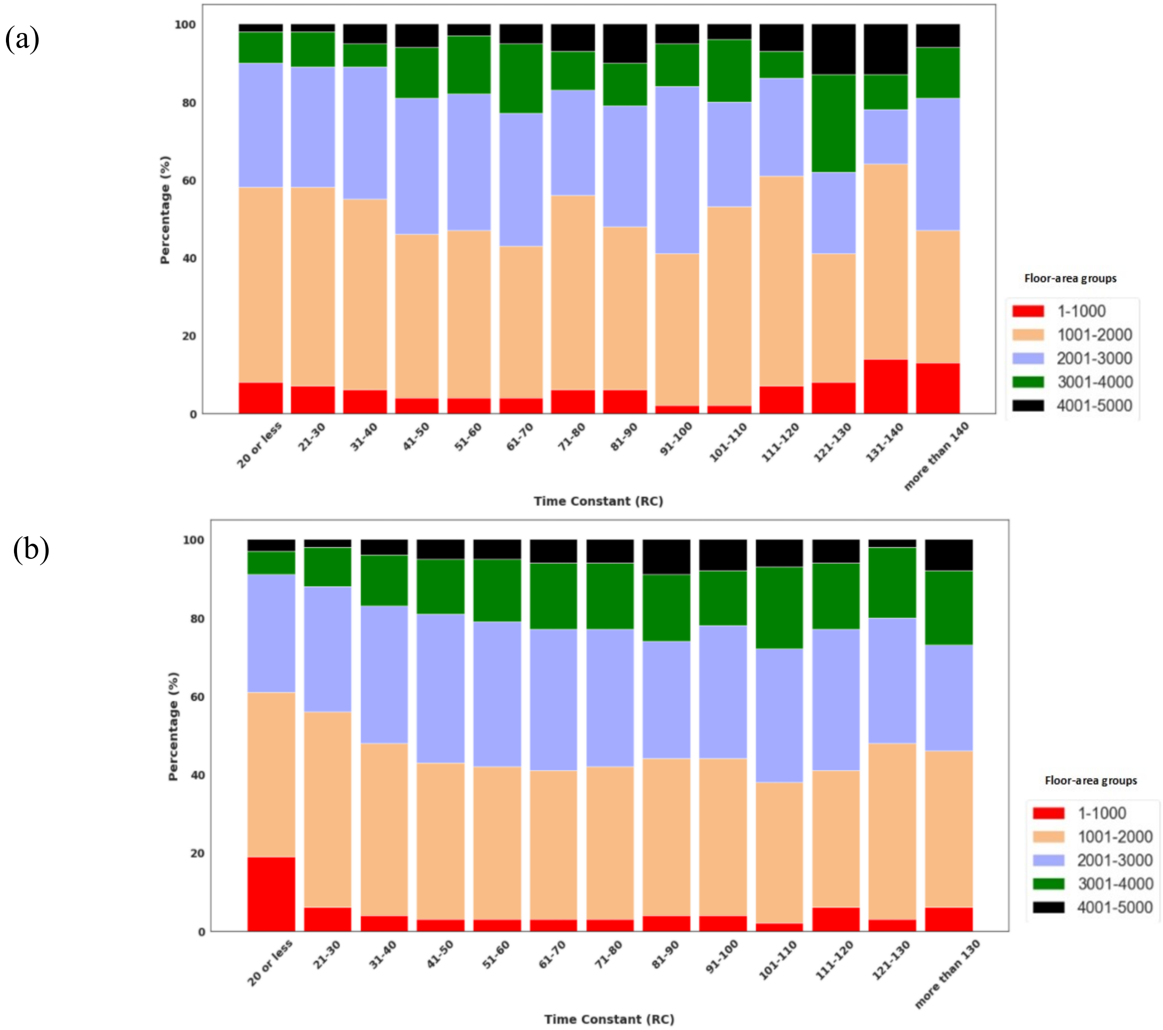


Figure 20: The distribution of each RC range category with respect to the floor areas for: (a) the decay curve RC-values and (b) the energy balance RC-values.

4.4.4 Investigating the Distribution of the RC-values with respect to ASHRAE Climate Zone

Figure 21 shows the distribution of the ASHRAE climate zones in each RC range category for the decay curve and the energy balance results. A decreasing trend can be seen in the percentages of the mixed zone (zone 4) while the percentages of the colder zones (zone 6 and 7) had an increasing trend especially in the energy balance results. These trends align with the code requirements of having higher thermal insulation for the houses in the colder climate zones which will lead to higher RC values than the houses in the warmer ones.

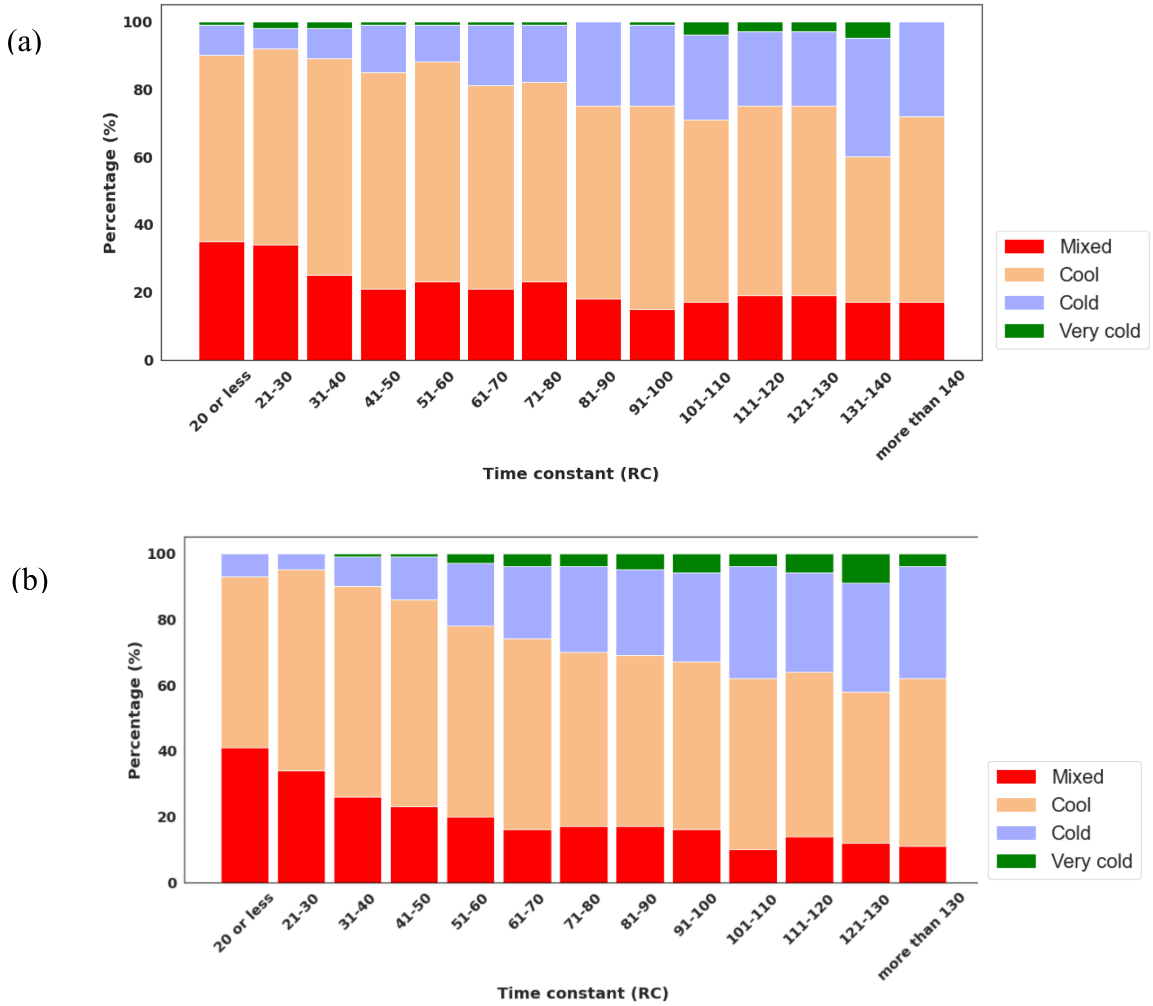


Figure 21: The distribution of each RC range category with respect to ASHRAE climate zone for: (a) the decay curve RC-values and (b) the energy balance RC-values.

4.5 The Two-step Multi-class Classification Model Results

To train the two-step multi-class classification model, four training subsets were used; Two without resampling the decay curve and the energy balance RC results, and two with resampled subsets. The Trained models were then evaluated using the validation subsets, which had 300 houses for the decay curve model and 4022 houses for the energy balance model. A summary of the evaluation results reported from the confusion matrixes and the KS test can be seen in Table 8, while the full confusion matrices can be seen in appendix 15- 18.

Recall that the two-step multi-class classification model aims to predict residential buildings' RC value within a range (15-hour bins) based on five attributes. Validating this model using the validation subset with the decay curve results indicated that 79% of the houses were either correctly classified into their 15-hour bin range or misclassified by only one bin. For the energy balance subset, this percentage decreased to 61% of houses.

Using the re-sampling method explained in section 3.5.4.2 to reduce imbalance within the training dataset resulted in decreasing the percentage of houses correctly classified into their 15-hour bin range or misclassified by only one bin to 71% for the decay curve subset, and 60% for the energy balance subset. However, the KS test results showed that resampling the training subsets significantly minimized the differences between the actual values and the predicted values distributions as shown in Table 8 and the confusion matrices shown in the appendix. These results suggest that re-sampling of the training dataset would be preferred for the proposed two-step multi-class classification model.

Table 8: The evaluation of the two-step multi-class classification model.

Training subsets	Classified into the same class	Misclassified by only one class	KS test results
Decay Curve			
Before resampling	31%	48%	0.25
After resampling	17%	54%	0.03
Energy Balance			
Before resampling	38%	23%	0.16
After resampling	27%	32%	0.05

Conclusion and Future Works

In this study, two methods were used to estimate the thermal time constant (RC value) for more than 60,000 houses across Canada and US using ecobee smart thermostat dataset. The results from the decay curve method reported a median of 52 hours with 90% of the houses having RC values ranging from 11 to 80 hours. The energy balance results reported a median of 51 hours with 90% of the houses having RC values ranging from 11 to 110 hours. A positive linear correlation was found between the RC values obtained from the two methods. The RC values were then used to estimate the lower and the upper bound R-values for each house and compare it to benchmark R-values estimated from the latest version of the ASHRAE 90.2 code requirements. The estimated bounds for the R-values were lower than the ASHRAE-compliant R-values which can be due to the poor craftmanship or that the older houses followed less stringent energy codes.

Furthermore, the estimated RC values were analyzed with respect to the building's attributes. The results showed higher RC medians for larger houses and for the houses in the colder climate zones, while a decreasing trend was seen in the RC medians of smaller buildings, especially row houses and apartments. Finally, a two-step multi-class classification model was developed to predict the RC range for the houses based on their attributes, whose results suggested that re-sampling of the training datasets to reduce imbalance can minimize the differences between the distributions of actual and predicted values. The classification models were also used to rank the importance of the building's attributes on the predictions of the RC range. The floor area and ASHRAE climate zone had the highest influence on RC range predictions based on the decay curve model, while the energy balance model ranked the ASHRAE climate zone and building age as the highest attributes. The building style had the lowest impact on the RC range prediction based on the decay curve and the energy balance models.

By using a large-scale dataset, the results from this study can be very beneficial to prioritize existing buildings for retrofits as well as providing inputs for urban-scale simulations and generating renovation scenarios for different buildings clusters. However, they are subjected to some limitations and require further development. In the ecobee dataset, which was the core of this study, many building attributes such as age, floor area, and number of floors were user-based inputs which are prone to errors. This limitation was addressed in the data cleaning process and identified the outliers with respect to each reported building's attribute.

Another limitation of the used dataset is that the sample of buildings from which data was collected may not be representative of the entire building stock, despite its large size, given the relatively high price point of smart thermostats. In this respect, more data are required from the wider building stock to enhance the generalization and the accuracy of the results. Furthermore, the outdoor temperature readings in the dataset were approximated from nearby weather stations which could differ from microclimates at the exact building site. Since these measurements were used to estimate the RC value for each house, the proposed methodology to develop a model to predict the house thermal performance based on its attributes focused on predicting the RC ranges rather than the exact value.

In terms of the potential limitation of the decay curve and the energy balance methods, these methods assume that heat gains were attributed to the heating system only. Although specific periods were filtered out to ensure solar and occupant heat gains were minimized, internal heat gains may have still influenced recorded temperatures used to fit each model. To overcome this limitation, more detailed dataset is required which include the solar radiation and occupancy schedule. This can help to incorporate the solar and occupant heat gains in the estimation of the building’s thermal performance process.

A combination between the thermostat datasets and the energy usage data can also be discussed in future work. This combination could help estimating the building thermal capacitance (C) and the thermal resistance (R-value) as independent values instead of the RC values. Moreover, further research will also focus on exploring other data-mining techniques and approaches other than the proposed in this study to increase the accuracy of identifying the correlation between the building’s attribute and the thermal performance which could help improve the performance of the classification model.

References

[1] Energy Information Administration - EIA. 2019. n.d. Accessed August 13, 2020.

https://www.eia.gov/todayinenergy/detail.php?id=41433.

[2] United Nations Environment, 2017. Global Status Report.

https://www.worldgbc.org/sites/default/files/UNEP%20188_GABC_en%20%28web%29.pdf

[3] Biddulph, P., Gori, V., Elwell, C. A., Scott, C., Rye, C., Lowe, R., & Oreszczyn, T. (2014). Inferring the thermal resistance and effective thermal mass of a wall using frequent temperature and heat flux measurements. *Energy & Buildings*, *78*, 10–16. https://doi.org/10.1016/j.enbuild.2014.04.004

[4] Aznar, F., Echarri, V., Rizo, C., & Rizo, R. (2018). Modelling the thermal behaviour of a building facade using deep learning. *Plos One*, *13*(12). https://doi.org/10.1371/journal.pone.0207616

[5] Li, S. et al. (2014) "Energy Modeling of Photovoltaic Thermal Systems with Corrugated Unglazed Transpired Solar Collectors - Part 1: Model Development and Validation," Solar Energy, 102, pp. 282–296. doi: 10.1016/j.solener.2013.12.040.

[6] Notton, G. et al. (2013) "New Patented Solar Thermal Concept for High Building Integration: Test and Modeling," Energy Procedia, 42, pp. 43–52. doi: 10.1016/j.egypro.2013.11.004.

[7] Maurer, C. and Kuhn, T. E. (2012) "Variable g Value of Transparent Façade Collectors," Energy & Buildings, 51, pp. 177–184. doi: 10.1016/j.enbuild.2012.05.011.

[8] Motte, F. et al. (2013) "Design and Modelling of a New Patented Thermal Solar Collector with High Building Integration," Applied Energy, 102, pp. 631–639. doi: 10.1016/j.apenergy.2012.08.012.

[9] Arteconi, A., Hewitt, N. J. and Polonara, F. (2013) "Domestic Demand-Side Management (dsm): Role of Heat Pumps and Thermal Energy Storage (tes) Systems," Applied Thermal Engineering, 51(1-2), pp. 155–165. doi: 10.1016/j.applthermaleng.2012.09.023.

[10] Bozkaya, B. and Zeiler, W. (2020) "The Energy Efficient Use of an Air Handling Unit for Balancing an Aquifer Thermal Energy Storage System," Renewable Energy, 146, pp. 1932–1942.

[11] Wang, K. et al. (2018) "Thermal Energy Storage Tank Sizing for Biomass Boiler Heating Systems Using Process Dynamic Simulation," Energy & Buildings, 175, pp. 199–207. doi: 10.1016/j.enbuild.2018.07.023.

[12] Satyavada, H. and Baldi, S. (2018) "Monitoring Energy Efficiency of Condensing Boilers Via Hybrid First-Principle Modelling and Estimation," Energy, 142.

[13] Li, X., & Wen, J. (2014). Review of building energy modeling for control and operation. *Renewable and Sustainable Energy Reviews*, *37*, 517–537. https://doi.org/10.1016/j.rser.2014.05.056

[14] Wang, S., Yan, C., & Xiao, F. (2012). Quantitative energy performance assessment methods

for existing buildings. *Energy & Buildings*, *55*, 873–888. https://doi.org/10.1016/j.enbuild.2012.08.037

[15] Ayres, J. M., & Stamper, E. (1995). Historical development of building energy calculations. *ASHRAE Journal*, *37*(2).

[16] Mohandes, S. R., Zhang, X., & Mahdiyar, A. (2019). A comprehensive review on the application of artificial neural networks in building energy analysis. *Neurocomputing*, *340*, 55–75. https://doi.org/10.1016/j.neucom.2019.02.040

[17] Lundin, M., Andersson, S., & Östin Ronny. (2004). Development and validation of a method aimed at estimating building performance parameters. *Energy & Buildings*, *36*(9), 905–914. https://doi.org/10.1016/j.enbuild.2004.02.005

[18] Kalogirou, S. A., Neocleous, C. C., & Schizas, C. N. (1997). Building heating load estimation using artificial neural networks. In *Proceedings of the 17th international conference on Parallel architectures and compilation techniques* (Vol. 8, p. 14).

[19] Xu, X., Wang, W., Hong, T., & Chen, J. (2019). Incorporating machine learning with building network analysis to predict multi-building energy use. *Energy & Buildings*, *186*, 80–97. https://doi.org/10.1016/j.enbuild.2019.01.002

[20] Jiménez, M. J., Madsen, H., & Andersen, K. K. (2008). Identification of the main thermal characteristics of building components using matlab. *Building and Environment*, *43*(2), 170–180. https://doi.org/10.1016/j.buildenv.2006.10.030

[21] Yun, K., Luck, R., Mago, P. J., & Cho, H. (2012). Building hourly thermal load prediction using an indexed arx model. *Energy & Buildings*, *54*, 225–233. https://doi.org/10.1016/j.enbuild.2012.08.007

[22] Francois, A., Ibos, L., Feuillet, V., & Meulemans, J. (2020). Estimation of the thermal resistance of a building wall with inverse techniques based on rapid active in situ measurements and white-box or arx black-box models. *Energy and Buildings*, *226*. https://doi.org/10.1016/j.enbuild.2020.110346

[23] Kontokosta, C. E., & Tull, C. (2017). A data-driven predictive model of city-scale energy use in buildings. *Applied Energy*, *197*, 303–317.

[24] Ma, J., Qin, J., Salsbury, T., & Xu, P. (2012). Demand reduction in building energy systems based on economic model predictive control. *Chemical Engineering Science*, *67*(1), 92–100. https://doi.org/10.1016/j.ces.2011.07.052

[25] Gassar, A. A. A. and Cha, S. H. (2020) “Energy Prediction Techniques for Large-Scale Buildings Towards a Sustainable Built Environment: A Review,” Energy & Buildings, 224. doi: 10.1016/j.enbuild.2020.110238

[26] Gouda, M. M., Danaher, S., & Underwood, C. P. (2002). Building thermal model reduction using nonlinear constrained optimization. *Building and Environment*, *37*(12), 1255–1265. https://doi.org/10.1016/S0360-1323(01)00121-4

[27] Reynders, G., Diriken, J., & Saelens, D. (2014). Quality of grey-box models and identified parameters as function of the accuracy of input and observation signals. *Energy and Buildings*, *82*, 263-274.

[28] Ferracuti, F., Fonti, A., Ciabattoni, L., Pizzuti, S., Arteconi, A., Helsen, L., & Comodi, G. (2017). Data-driven models for short-term thermal behaviour prediction in real buildings. *Applied Energy*, *204*, 1375–1387.

[29] Baasch, G., Wicikowski, A., Faure, G., & Evins, R. (2019, November). Comparing gray box methods to derive building properties from smart thermostat data. In *Proceedings of the 6th ACM international conference on systems for energy-efficient buildings, cities, and transportation* (pp. 223-232).

[30] "Heating balance point temperatures https://www.caltrack.org/" .

[31] Tabatabaei, S., van der Ham, W., C. A. Klein, M., & Treur, J. (2017). A data analysis technique to estimate the thermal characteristics of a house. *Energies*, *10*(9), 1358–1358. https://doi.org/10.3390/en10091358

[32] van der Ham, W., Klein, M., Tabatabaei, S. A., Thilakarathne, D., & Treur, J. (2016). Methods for a smart thermostat to estimate the characteristics of a house based on sensor data with varying extent of completeness. *Energy Procedia*, *95*, 467–474.

[33] John, C., Vallianos, C., Candanedo, J., & Athienitis, A. (2018). Estimating time constants for over 10,000 residential buildings in North America: towards a statistical characterization of thermal dynamics.

[34] Hossain, M.M., Zhang, T. and Ardakanian, O., (2020). Identifying Grey-box Thermal Models with Bayesian Neural Networks. arXiv preprint arXiv:2009.05889.

[35] Ecobee Inc., "Ecobee Donate Your Data Program," (2019). [Online]. Available: https://www.ecobee.com/donateyourdata/.

[36] Zhu, N. et al. (2010). A Simplified Dynamic Model of Building Structures Integrated with Shaped-Stabilized Phase Change Materials, International Journal of Thermal Sciences, 49(9), pp. 1722–1731. doi: 10.1016/j.ijthermalsci.2010.03.020.

[37] Hedegaard, R. E. and Petersen, S. (2017). Evaluation of Grey-Box Model Parameter Estimates Intended for Thermal Characterization of Buildings," Energy Procedia, 132, pp. 982–987. doi: 10.1016/j.egypro.2017.09.692.

[38] Bacher, P. and Madsen, H. (2011) Identifying Suitable Models for the Heat Dynamics of Buildings. Energy & Buildings, 43(7), pp. 1511–1522. doi: 10.1016/j.enbuild.2011.02.005.

[39] Brastein, O. M. et al. (2018) Parameter Estimation for Grey-Box Models of Building Thermal Behaviour. Energy & Buildings, 169, pp. 58–68. doi: 10.1016/j.enbuild.2018.03.057.

[40] Blum, D. H. et al. (2019) Practical Factors of Envelope Model Setup and Their Effects on the Performance of Model Predictive Control for Building Heating, Ventilating, and Air Conditioning Systems. Applied Energy, 236, pp. 410–425. doi: 10.1016/j.apenergy.2018.11.093.

[41] Hollick, F. P., Gori, V. and Elwell, C. A. (2020). Thermal Performance of Occupied Homes: A Dynamic Grey-Box Method Accounting for Solar Gains. Energy & Buildings, 208. doi: 10.1016/j.enbuild.2019.109669.

[42] Dunn, O.J. (1964). Multiple Comparisons Using Rank Sums. Technometrics, vol. 6, no. 3, pp. 241–252.

[43] Freund, Y., & Shapire, R., (1995). A decision-theoretic generalization of on-line learning and application to boosting. Proceedings of the Second European Conference on Computaional Learning Theory. *Computer and System Sciences*, *55*, 119–139.

[44] Nahlik, M. J. et al. (2017). Building Thermal Performance, Extreme Heat, and Climate Change. Journal of infrastructure systems, 23(3), pp. 04016043–04016043

Appendix

Appendix 1: Descriptive statistics of the RC values estimated from the energy balance and the decay curve method for different building-style.

Energy Balance Results				
Building styles	Mean	Standard Deviation	Median	Number of Houses
Apartment	64.46	31.46	55.82	751
Row House	59.79	23.25	79.33	2014
Semi-Detached	60.85	22.26	70.10	592
Detached	54.00	20.60	58.41	13,548
Others	53.68	21.53	52.19	5016
Total				**21,921**
Decay Curve Results				
Building styles	Mean	Standard Deviation	Median	Number of Houses
Apartment	68.60	53.45	79.09	101
Row House	64.78	37.21	68.16	222
Semi-Detached	72.43	49.34	61.43	69
Detached	56.51	31.97	58.11	1135
Others	54.92	32.75	55.73	585
Total				**2112**

Appendix 2: Descriptive statistics of the RC values estimated from the energy balance and the decay curve method for different floor-area.

Energy Balance Results				
Floor area	Mean	Standard Deviation	Median	Number of Houses
1-1000	53.64	25.95	48.87	739
1001-2000	53.87	21.68	49.56	7982
2001-3000	54.42	20.68	50.37	7046
3001-4000	57.17	21.19	53.14	2880
4001-5000	59.28	22.01	54.68	971
Total				**19,618**
Decay Curve Results				
Floor area	Mean	Standard Deviation	Median	Number of Houses

1-1000	61.70	50.20	49.04	102
1001-2000	56.54	34.09	50.17	824
2001-3000	57.08	32.67	51.18	600
3001-4000	62.30	32.46	55.91	215
4001-5000	66.74	34.47	62.33	91
Total				**1832**

Appendix 3: Descriptive statistics of the RC values estimated from the energy balance and the decay curve method for different building-age.

Energy Balance Results

Building age	Mean	Standard Deviation	Median	Number of Houses
1-10	61.02	23.78	54.67	4252
11-20	55.12	21.52	54.24	3838
21-30	52.53	17.52	54.47	2370
31-40	51.76	18.71	46.81	1914
41-50	51.53	19.21	47.21	1520
51-60	51.11	18.39	46.27	1405
61-70	51.74	18.87	50.53	896
71-80	50.88	19.41	51.51	411
81-90	50.86	20.44	50.89	654
91-100	51.81	21.73	58.24	398
101-110	53.09	21.38	65.38	254
111-120	52.60	24.88	65.73	407
Total				**18,319**

Decay Curve Results

Building age	Mean	Standard Deviation	Median	Number of Houses
1-10	58.89	33.12	56.91	354
11-20	58.18	32.44	51.15	365
21-30	61.65	39.07	49.18	202
31-40	53.26	30.63	47.78	166
41-50	53.60	35.55	47.75	129
51-60	56.74	36.35	48.66	138
61-70	52.97	26.62	48.19	97
71-80	54.78	21.48	47.40	47
81-90	57.32	33.23	45.995	81

91-100	58.18	28.20	47.06	45
101-110	67.98	27.03	49.75	29
111-120	68.41	44.93	47.097	43
Total				**1696**

Appendix 4: Descriptive statistics of the RC values estimated from the energy balance and the decay curve method for different ASHRAE climate zones.

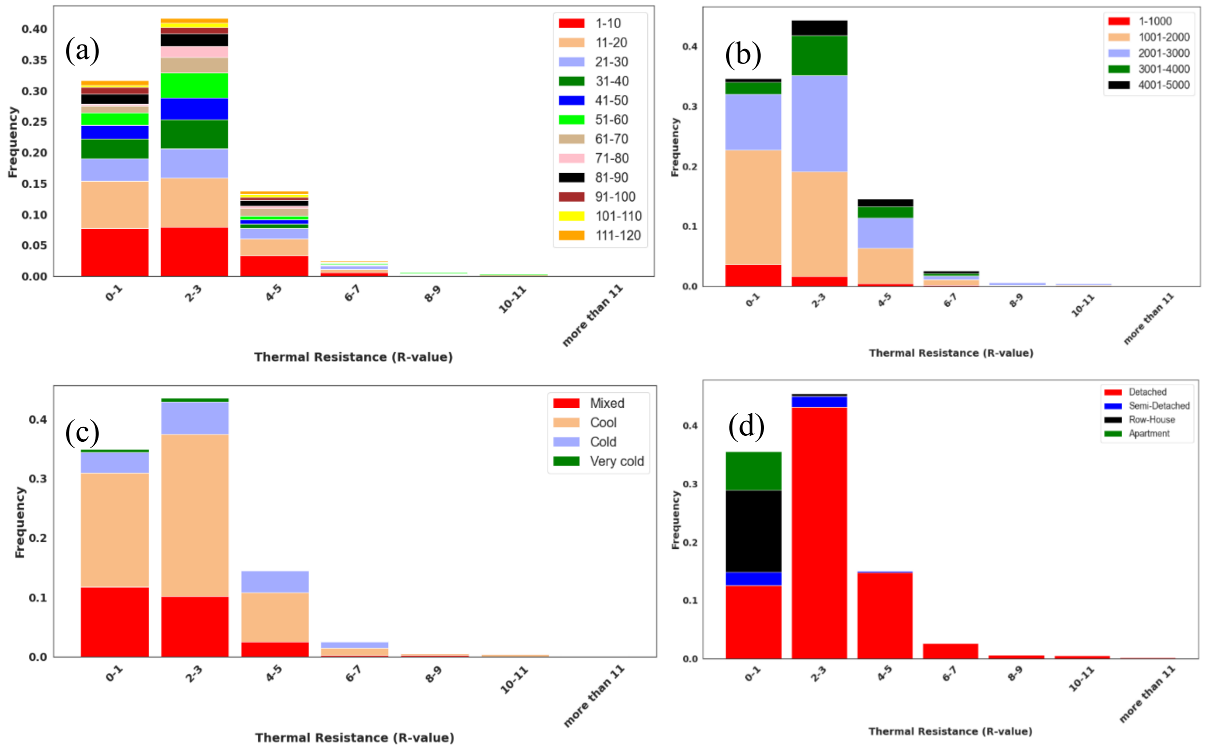
Energy Balance Results

ASHRAE Climate Zone	Mean	Standard Deviation	Median	Number of Houses
Mixed	50.13	19.62	46.42	4410
Cool	53.22	20.53	49.18	12247
Cold	62.79	23.93	58.28	3419
Very cold	67.40	23.84	63.31	481
Total				**20,559**

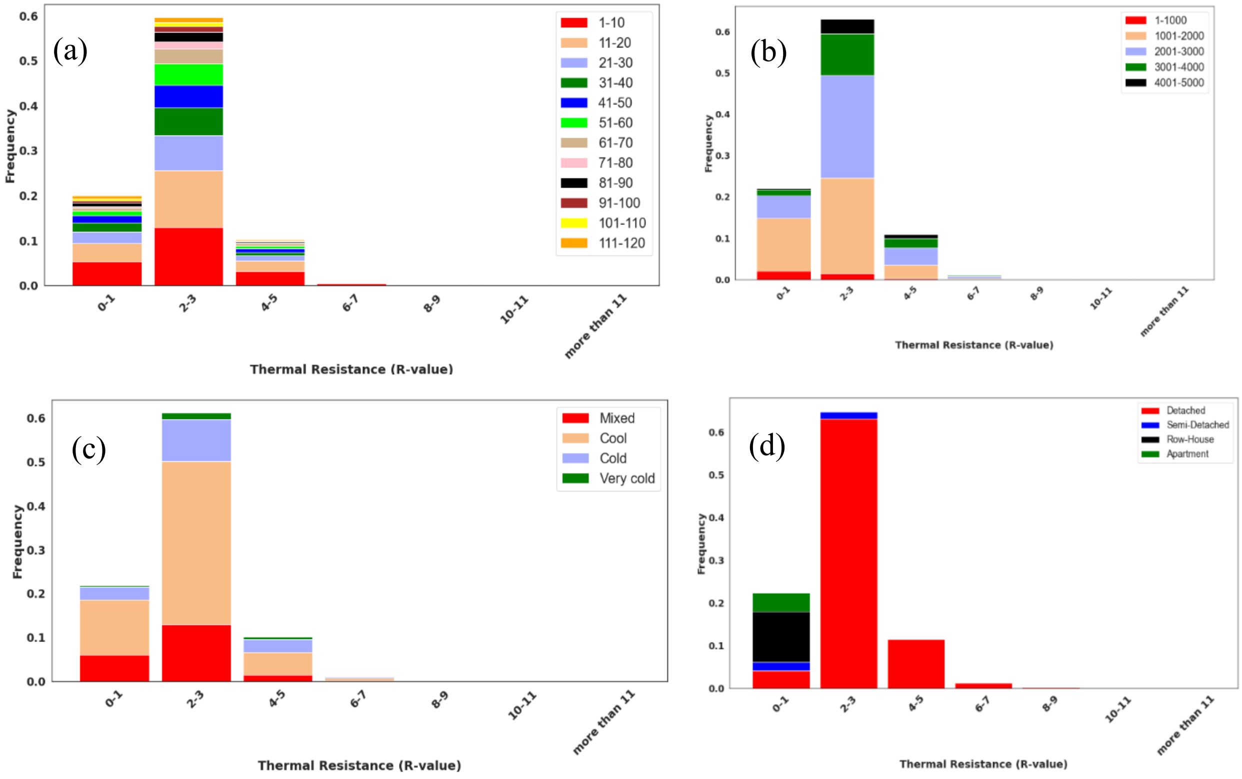
Decay Curve Results

ASHRAE Climate Zone	Mean	Standard Deviation	Median	Number of Houses
Mixed	51.92	33.76	46.42	481
Cool	57.47	34.43	51.33	1225
Cold	71.02	37.62	64.43	292
Very cold	57.48	33.91	47.3	29
Total				**2027**

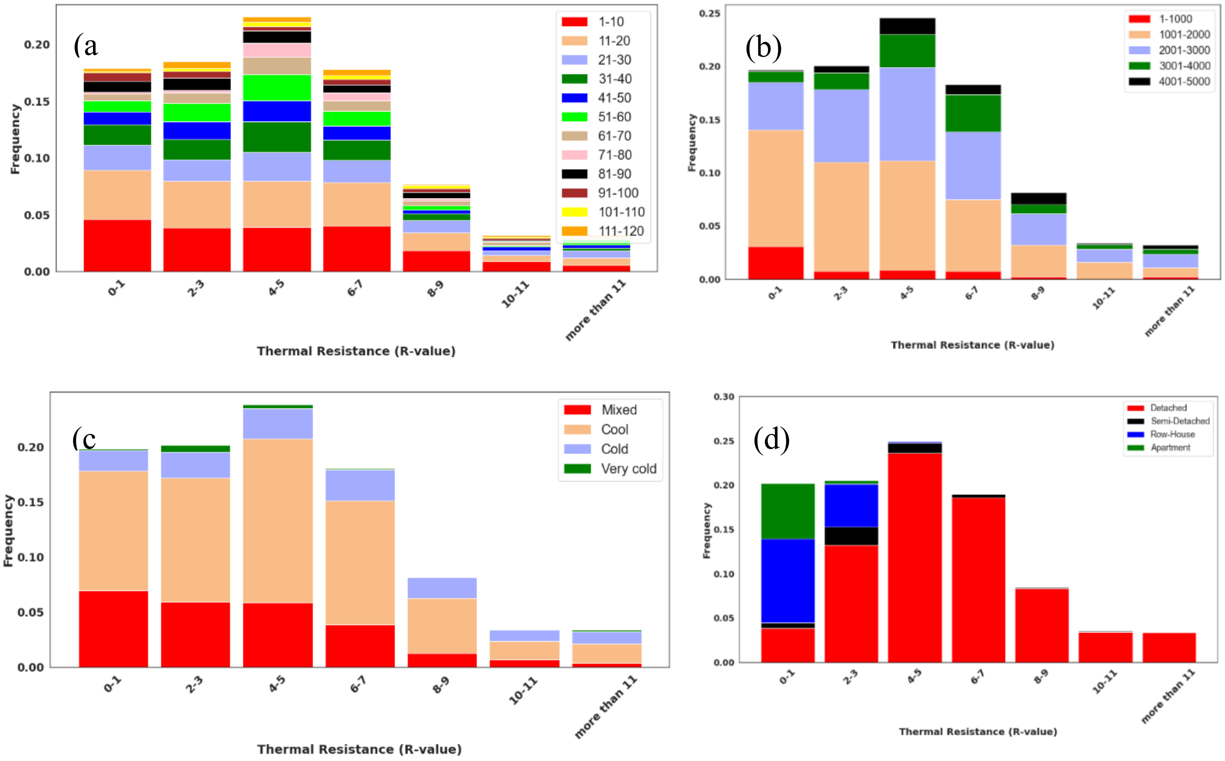
Appendix 5: The distribution of the estimated minimum R-values from the decay curve model with respect to the building's attributes.



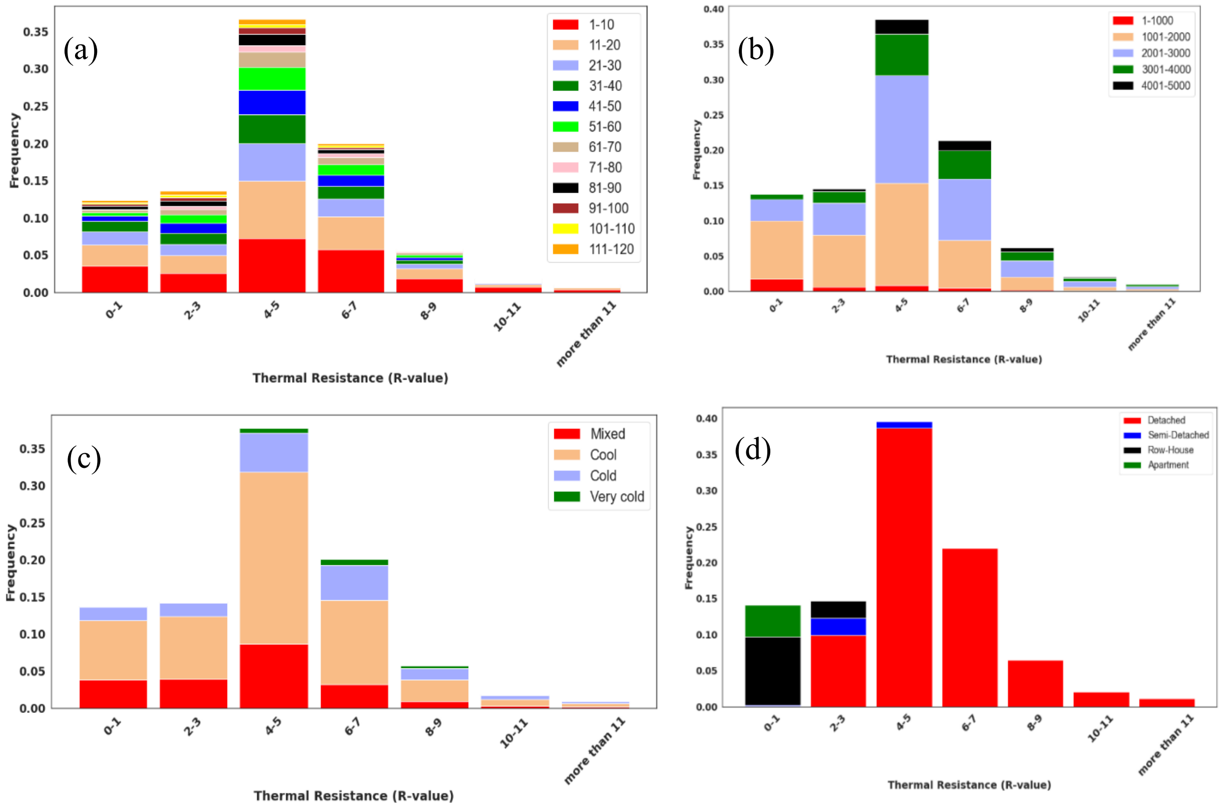
Appendix 6: The distribution of the estimated minimum R-values from the energy balance model with respect to the building's attributes.



Appendix 7: The distribution of the estimated maximum R-values from the decay curve model with respect to the building's attributes.



Appendix 8: The distribution of the estimated maximum R-values from the energy balance model with respect to the building's attributes.



Appendix 9: The results of the Kruskal Wallis test for the energy balance and the decay curve RC-values.

Kruskal Wallis test Results				
Energy balance Results				
Attributes	H-value	P-value	Degree of freedom	Chi-squares
Building- styles	296.02	$1.58\text{x}10^{-63}$	4	9.488
Building-age	656.81	$9.67\text{x}10^{-134}$	11	19.675
Floor area	144.92	$2.5\text{x}10^{-30}$	4	9.488
ASHRAE' climate zone	1094.86	$4.75\text{x}10^{-237}$	3	7.815
Number of Floors	20.20	$4.099\text{x}10^{-05}$	2	5.991
Decay Curve Results				
Attributes	H-value	P-value	Degree of freedom	Chi-squares
Building- styles	20.21	0.0005	4	9.488
Building-age	21.45	0.03	11	19.675

Floor area	17.755	0.0014	4	9.488
ASHRAE' climate zone	69.39	5.75x10^{-15}	3	7.815
Number of Floors	148.625	5.33x10^{-33}	2	5.991

Appendix 10: The p-value results of the Dunn pair-comparison test for the energy balance and the decay curve RC-values for different building-style.

Dunn pair-comparison test

Decay Curve Results

	Apartment	Row-House	Semi-Detached	Detached	Other
Apartment	-1	0.99996	0.99996	0.99996	0.41327
Row-House	0.99996	-1	0.99996	0.03238	0.00265
Semi-Detached	0.99996	0.99996	-1	0.12016	0.03120
Detached	0.99996	0.03238	0.12015	-1	0.66231
Other	0.41327	0.00265	0.03120	0.66231	-1

Energy Balance

	Apartment	Row-House	Semi-Detached	Detached	Other
Apartment	-1	0.99996	0.99996	0.99996	0.41327
Row-House	0.99996	-1	0.99996	0.03238	0.00265
Semi-Detached	0.99996	0.99996	-1	0.12016	0.03120
Detached	0.99996	0.03238	0.12015	-1	0.66231
Other	0.41327	0.00265	0.03120	0.66231	-1

Appendix 11: The p-value results of the Dunn pair-comparison test for the energy balance and the decay curve RC-values for different floor-area.

Dunn pair-comparison test

Decay Curve Results

	1-1000	1001-2000	2001-3000	3001-4000	4001-5000
1-1000	-1.00	1.00	1.00	0.35	0.15
1001-2000	1.00	-1.00	1.00	0.03	0.03
2001-3000	1.00	1.00	-1.00	0.12	0.08
3001-4000	0.35	0.03	0.12	-1.00	1.00
4001-5000	0.15	0.03	0.08	1.00	-1.00

Energy Balance

	1-1000	1001-2000	2001-3000	3001-4000	4001-5000

1-1000	-1	0.168554756	0.019	3.36E-08	7.32E-12
1001-2000	0.17	-1	0.01	2.26E-17	2.11E-17
2001-3000	0.01853219	0.01	-1	4.91E-09	7.90E-12
3001-4000	3.36E-08	2.26E-17	4.91E-09	-1	0.014
4001-5000	7.32E-12	2.11E-17	7.90E-12	0.014	-1

Appendix 12: The p-value results of the Dunn pair-comparison test for the energy balance and the decay curve RC-values for different building-age.

Dunn Pair-comparison Test Results

Decay Curve Results

	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100	101-110	111-120
1-10	-1	1	1	1	1	1	1	1	1	1	1	1
11-20	1	-1	1	1	1	1	1	1	1	1	1	1
21-30	1	1	-1	1	1	1	1	1	1	1	1	1
31-40	1	1	1	-1	1	1	1	1	1	1	0.38	1
41-50	1	1	1	1	-1	1	1	1	1	1	0.27	0.75
51-60	1	1	1	1	1	-1	1	1	1	1	0.95	1
61-70	1	1	1	1	1	1	-1	1	1	1	1	1
71-80	1	1	1	1	1	1	1	-1	1	1	1	1
81-90	1	1	1	1	1	1	1	1	-1	1	1	1
91-100	1	1	1	1	1	1	1	1	1	-1	1	1
101-110	1	1	1	0.38	0.27	0.95	1	1	1	1	-1	1
111-120	1	1	1	1	0.75	1	1	1	1	1	1	-1

Energy Balance

	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100	101-110	111-120
1-10	-1	8.68E-35	5.81E-50	4.75E-60	1.16E-55	3.71E-52	9.57E-33	2.29E-21	1.31E-35	5.61E-20	1.16E-08	1.64E-19
11-20	8.68E-35	-1	0.002	2.13E-08	5.88E-09	2.54E-08	0.0003	0.0004	8.66E-08	0.001	1	0.003
21-30	5.81E-50	0.002	-1	1	0.3	0.4	1	0.9	0.04	1	1	1
31-40	4.75E-60	2.13E-08	1	-1	1	1	1	1	1	1	1	1
41-50	1.16E-55	5.88E-09	0.3	1	-1	1	1	1	1	1	1	1
51-60	3.71E-52	2.54E-08	0.4	1	1	-1	1	1	1	1	1	1
61-70	9.57E-33	0.0003	1	1	1	1	-1	1	1	1	1	1
71-80	2.29E-21	0.0004	0.9	1	1	1	1	-1	1	1	1	1
81-90	1.31E-35	8.66E-08	0.04	1	1	1	1	1	-1	1	1	1

91-100	5.61E-20	0.001	1	1	1	1	1	1	1	-1	1	1
101-110	1.16E-08	1	1	1	1	1	1	1	1	1	-1	1
111-120	1.64E-19	0.002614	1	1	1	1	1	1	1	1	1	-1

Appendix 13: The p-value results of the Dunn pair-comparison test for the energy balance and the decay curve RC-values for ASHREA's climate zones.

Dunn pair-comparison test

Decay Curve Results

	mixed	cool	cold	very cold
mixed	-1	0.002	1.18E-15	1
cool	0.002	-1	1.19E-09	1
cold	1.18E-15	1.19E-09	-1	0.095
very cold	1	1	0.095	-1

Energy Balance

	mixed	cool	cold	very cold
mixed	-1	8.86E-21	1.13E-178	2.72E-74
cool	8.86E-21	-1	1.06E-136	3.12E-52
cold	1.13E-178	1.06E-136	-1	1.77E-05
very cold	2.72E-74	3.12E-52	1.77E-05	-1

Appendix 14: Number of the houses in each class in the training datasets before and after clustering.

Decay Curve training dataset distribution

	0-15	16-30	31-45	46-60	61-70	71-90	91-105	106-120	121-135	136-150	More than 150
Before resampling	63	156	237	243	187	100	75	21	27	6	22
After resampling	243	243	243	243	243	243	243	243	243	243	243

Energy Balance training dataset distribution

	0-15	16-30	31-45	46-60	61-70	71-90	91-105	106-120	121-135	136-150	More than 150
Before resampling	16	712	4013	3930	1983	811	310	148	70	23	47
After resampling	100	500	500	500	500	500	500	100	100	100	100

Appendix 15: Confusion matrix for the decay curve model without resample the training subset.

		predicted value											
		0-15	16-30	31-45	46-60	61-75	76-90	91-105	106-120	121-135	136-150	more than 150	Total number of houses
Actual value	0-15	0.11	0.00	0.53	0.26	0.05	0.05	0.00	0.00	0.00	0.00	0.00	19
	16-30	0.02	0.15	0.49	0.22	0.05	0.00	0.05	0.02	0.00	0.00	0.00	41
	31-45	0.01	0.03	0.70	0.22	0.03	0.00	0.00	0.00	0.00	0.00	0.00	67
	46-60	0.02	0.09	0.41	0.41	0.06	0.00	0.00	0.02	0.00	0.00	0.00	66
	61-75	0.00	0.07	0.41	0.32	0.14	0.02	0.02	0.02	0.00	0.00	0.00	44
	76-90	0.00	0.06	0.39	0.32	0.00	0.13	0.06	0.03	0.00	0.00	0.00	31
	91-105	0.00	0.00	0.27	0.27	0.47	0.00	0.00	0.00	0.00	0.00	0.00	15
	106-120	0.25	0.00	0.50	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4
	121-135	0.00	0.25	0.50	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4
	136-150	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1

	more than 150	0.00	0.13	0.75	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	8

Appendix 16: Confusion matrix for the decay curve model After resample the training subset.

		predicted value											
		0-15	**16-30**	**31-45**	**46-60**	**61-75**	**76-90**	**91-105**	**106-120**	**121-135**	**136-150**	**more than 150**	**Total number of houses**
Actual value	**0-15**	0.05	0.10	0.14	0.38	0.05	0.05	0.10	0.05	0.05	0.05	0.00	19
	16-30	0.06	0.15	0.31	0.15	0.17	0.08	0.04	0.02	0.00	0.00	0.02	41
	31-45	0.04	0.23	0.21	0.22	0.15	0.07	0.02	0.01	0.01	0.00	0.02	67
	46-60	0.04	0.18	0.16	0.31	0.13	0.13	0.01	0.01	0.01	0.01	0.00	66
	61-75	0.09	0.12	0.26	0.23	0.11	0.11	0.05	0.04	0.00	0.00	0.00	44
	76-90	0.10	0.18	0.13	0.23	0.18	0.13	0.00	0.03	0.03	0.03	0.00	31
	91-105	0.06	0.06	0.00	0.28	0.39	0.06	0.00	0.00	0.06	0.06	0.06	15
	106-120	0.00	0.36	0.18	0.09	0.18	0.00	0.00	0.09	0.00	0.00	0.09	4
	121-135	0.20	0.20	0.00	0.40	0.20	0.00	0.00	0.00	0.00	0.00	0.00	4

	136-150	0.00	0.00	0.00	0.50	0.50	0.00	0.00	0.00	0.00	0.00	0.00	1
	more than 150	0.00	0.00	0.44	0.33	0.11	0.00	0.00	0.00	0.00	0.00	0.11	8

Appendix 17: Confusion matrix for the energy balance model without resample the training subset.

		predicted value											
		0-15	**16-30**	**31-45**	**46-60**	**61-75**	**76-90**	**91-105**	**106-120**	**121-135**	**136-150**	**more than 150**	**Total number of houses**
Actual value	**0-15**	0.20	0.00	0.40	0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5
	16-30	0.10	0.01	0.72	0.12	0.01	0.00	0.01	0.03	0.00	0.00	0.00	249
	31-45	0.06	0.01	0.67	0.22	0.01	0.00	0.00	0.02	0.00	0.00	0.00	1314
	46-60	0.04	0.01	0.60	0.31	0.01	0.00	0.00	0.01	0.00	0.00	0.00	1301
	61-75	0.33	0.07	0.26	0.20	0.09	0.01	0.01	0.01	0.00	0.00	0.00	676
	76-90	0.03	0.00	0.50	0.41	0.02	0.00	0.00	0.02	0.00	0.00	0.00	271
	91-105	0.02	0.00	0.48	0.39	0.06	0.00	0.03	0.01	0.00	0.00	0.00	121
	106-120	0.02	0.00	0.40	0.48	0.05	0.00	0.00	0.05	0.00	0.00	0.00	51

121-135	0.00	0.00	0.33	0.50	0.08	0.00	0.00	0.08	0.00	0.00	0.00	15
136-150	0.00	0.00	0.64	0.36	0.00	0.00	0.00	0.00	0.00	0.00	0.00	13
more than 150	0.00	0.00	0.75	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	6

Appendix 18: Confusion matrix for the energy balance model after resample the training subset.

		predicted value											
		0-15	**16-30**	**31-45**	**46-60**	**61-75**	**76-90**	**91-105**	**106-120**	**121-135**	**136-150**	**more than 150**	**Total number of houses**
Actual value	**0-15**	0.00	0.20	0.20	0.60	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5
	16-30	0.00	0.22	0.38	0.25	0.06	0.04	0.02	0.00	0.01	0.00	0.00	249
	31-45	0.00	0.13	0.31	0.33	0.15	0.04	0.03	0.01	0.00	0.00	0.00	1314
	46-60	0.00	0.11	0.26	0.33	0.20	0.04	0.03	0.01	0.01	0.00	0.00	1301
	61-75	0.00	0.08	0.22	0.34	0.22	0.06	0.02	0.01	0.02	0.01	0.00	676
	76-90	0.01	0.09	0.23	0.31	0.20	0.08	0.04	0.02	0.01	0.00	0.00	271
	91-105	0.00	0.09	0.20	0.31	0.24	0.09	0.02	0.01	0.02	0.00	0.01	121
	106-120	0.02	0.12	0.24	0.25	0.20	0.06	0.04	0.02	0.04	0.02	0.00	51
	121-135	0.00	0.00	0.20	0.27	0.27	0.13	0.00	0.00	0.13	0.00	0.00	15

	136-150	0.00	0.08	0.23	0.15	0.23	0.15	0.15	0.00	0.00	0.00	0.00	13
	more than 150	0.00	0.00	0.17	0.33	0.17	0.17	0.00	0.17	0.00	0.00	0.00	6