



**Titre:** Development of a Bottom-up White-box Residential Building Stock  
Title: Energy Model

**Auteur:** Adam Neale  
Author:

**Date:** 2021

**Type:** Mémoire ou thèse / Dissertation or Thesis

**Référence:** Neale, A. (2021). Development of a Bottom-up White-box Residential Building  
Stock Energy Model [Thèse de doctorat, Polytechnique Montréal]. PolyPublie.  
Citation: <https://publications.polymtl.ca/9973/>

 **Document en libre accès dans PolyPublie**  
Open Access document in PolyPublie

**URL de PolyPublie:** <https://publications.polymtl.ca/9973/>  
PolyPublie URL:

**Directeurs de  
recherche:** Michel Bernier, & Michaël Kummert  
Advisors:

**Programme:** PhD.  
Program:

**POLYTECHNIQUE MONTRÉAL**

affiliée à l'Université de Montréal

**Development of a bottom-up white-box residential building stock energy  
model**

**ADAM NEALE**

Département de génie mécanique

Thèse présentée en vue de l'obtention du diplôme de *Philosophiæ Doctor*

Génie mécanique

Décembre 2021

# **POLYTECHNIQUE MONTRÉAL**

affiliée à l'Université de Montréal

Cette thèse intitulée:

## **Development of a bottom-up white-box residential building stock energy model**

présentée par **Adam NEALE**

en vue de l'obtention du diplôme de *Philosophiæ Doctor*

a été dûment acceptée par le jury d'examen constitué de :

**Jean-Yves TRÉPANIÉ**, président

**Michel BERNIER**, membre et directeur de recherche

**Michaël KUMMERT**, membre et codirecteur de recherche

**Richard LABIB**, membre

**Ursula EICKER**, membre externe

## ACKNOWLEDGEMENTS

I would like to start by gratefully acknowledging the support of my two supervisors, professors Michel Bernier and Michaël Kummert. This thesis would not have been possible without their guidance and I am very grateful that they took a chance on an “experienced” student such as myself. It has been a distinct pleasure working with them. Merci Michel et Michaël.

I am grateful to the members of my thesis examination committee, professors Jean-Yves Trépanier, Richard Labib and Ursula Eicker, for agreeing to review my thesis.

Thank you to the *Institut de Valorisation des données (IVADO)* for providing a generous research grant that supported this work and for recognizing the need for improvements to the field of building stock energy modeling. I appreciate the contributions and feedback that the project partners provided to my work along the way.

Many people have supported me over the years, and it would be difficult to name them all. Family, friends, work colleagues, and in-laws have all contributed to bring me to the point where I am today, and my success is thanks to them. My parents, Roger and Patricia, and my brother William, deserve special mention for their role in making me who I am today.

My two children, Marguerite and Jack, gave me energy when I did not have any, distraction when I needed it (and when I did not) and contributed significantly to my work by listening to me ramble on about smart meter data over the course of several years. As my work grew, so did they, and I am very proud of who they have become. Thank you M & J.

Finally, and most importantly, I owe this work to my wife Danielle, whose love and support allowed me to pursue a dream that I put aside when we were younger. She is a source of inspiration for me, and I can only aspire to be as good a parent, partner and professor, as she is. Merci Danielle, de tout mon coeur.

## RÉSUMÉ

La modélisation énergétique d'un parc immobilier, ou *Building stock energy modeling* (BSEM), est le processus d'évaluation de la consommation d'énergie des bâtiments à grande échelle, comme la consommation énergétique à l'échelle d'une ville, d'une région ou d'un pays. L'industrie canadienne a actuellement un besoin urgent pour le développement d'outils additionnels de modélisation du parc immobilier afin d'évaluer des mesures d'efficacité énergétique, d'estimer la réduction des émissions de gaz à effet de serre, d'évaluer l'impact des technologies ainsi que de nombreuses autres applications. Cette thèse aborde certaines lacunes liées au développement et à l'utilisation des BSEM, tout en fournissant un nouveau BSEM pour un parc immobilier de maisons unifamiliales au Canada.

Bien qu'il existe de nombreux types de BSEM, le processus nécessite toujours des informations sur le parc immobilier afin de générer un modèle des bâtiments. L'information disponible pour un parc de bâtiments donné est limitée par la quantité de données mesurées et par la divulgation publique de ces informations. Une nouvelle méthode de caractérisation des différents paramètres d'un parc immobilier résidentiel, tels que la surface chauffée, la résistance thermique des murs et le nombre d'occupants, est présentée. Pour pallier l'absence de jeux de données publiques appropriés, un jeu de données de compteurs intelligents virtuels, *virtual smart meter* (VSM) *data*, de 200 000 maisons avec 21 paramètres de construction connus pour chaque maison est développé. Ce jeu de données est disponible en libre accès. Les paramètres sont sélectionnés pour leur influence estimée sur la consommation d'électricité et leur applicabilité au processus de caractérisation d'un parc immobilier. Le jeu de données VSM est utilisé pour appliquer la classification par apprentissage automatique supervisé afin de développer des modèles prédictifs (MP), qui sont basés sur une analyse discriminante linéaire. Les MP peuvent estimer avec précision les paramètres des bâtiments, tels que la surface chauffée, à partir de la consommation d'électricité. Avec un jeu de données de compteurs intelligents de taille suffisante, la distribution des paramètres dans le parc immobilier peut être établie, ce qui permet de comprendre la probabilité qu'une maison appartienne à une catégorie spécifique. Des lignes directrices sont fournies pour les futures études de classification qui souhaitent utiliser un jeu de données réelles de compteurs intelligents dans le but de générer un modèle prédictif ou pour ceux qui souhaitent effectuer une classification des données de compteurs intelligents.

Un nouveau modèle énergétique du parc immobilier de maisons unifamiliales est développé, appelé le *Québec Single-Family Building Stock Energy Model (QSFBSSEM)*. Le QSFBSSEM prédit avec précision la consommation d'énergie des maisons du Québec, au Canada. Environ 1,9 million de maisons unifamiliales sont représentés par le modèle. La prédiction est faite pour le chauffage des locaux, le chauffage de l'eau, l'éclairage, les appareils électroménagers et la climatisation pour des maisons individuelles et jumelées. L'électricité, le gaz naturel, le mazout et le bois sont les principales sources d'énergie pour les bâtiments résidentiels dans la province et la répartition de la consommation de chaque source est également estimée par le modèle du parc immobilier. Un échantillon de 200 000 maisons est modélisé pour représenter l'ensemble du parc immobilier. La taille de l'échantillon est un aspect important pour la prédiction énergétique d'un parc immobilier puisque des parties moins représentées du parc peuvent entraîner une déviation significative de la prédiction énergétique si la taille de l'échantillon est trop faible. Le QSFBSSEM peut être utilisé pour une variété d'applications et est appliqué à une étude de cas pour la comparaison des émissions de gaz à effet de serre (GES) entre différentes configurations de systèmes de chauffage pour le parc immobilier. L'étude de cas permet de démontrer que le QSFBSSEM peut quantifier l'impact sur la demande de pointe en électricité d'une conversion massive des systèmes de chauffage non-électriques.

## **ABSTRACT**

Building stock energy modeling (BSEM) is the process of evaluating the energy consumption of large-scale building energy simulation problems, such as the energy use at the city, regional, or national levels. There is a current urgent need in industry in Canada for the development of additional stock modeling tools, for evaluation of energy efficiency measures, estimation of greenhouse gas emissions reductions, technology impact evaluations, and many other applications. This thesis addresses some of the common limitations associated with the development and use of BSEM, while providing a new BSEM for a single-family dwelling stock in Canada.

While many types of BSEM exist, the process always requires information on the building stock in order to reproduce the buildings in the form of a model. Obtaining sufficient data depends on the public availability of information and measured data for the studied building stock. A new method of characterizing the different parameters of a residential building stock, such as the heated surface area, wall thermal resistance and the number of occupants, is presented. Due to the lack of appropriate public data sets, a virtual smart meter (VSM) data set of 200,000 homes with 21 known building parameters per dwelling is designed and presented as an open-source data set. The parameters are selected due to their estimated impact on the electricity consumption and the potential applicability for the characterization process of a building stock. The VSM data set is used to apply supervised machine learning classification to develop predictive models (PM), which are in turn developed with linear discriminant analysis. The PM can accurately estimate building parameters such as the heated surface area from electricity consumption. With a sufficiently sized smart meter data set, the distribution of parameters across the building stock can be established, providing an understanding of the probability of a house belonging to a particular category. Guidelines are provided for future classification studies wishing to develop a real smart meter data set for the purpose of predictive model development or for those wishing to perform classification of smart meter data.

A new single-family dwelling building stock energy model is developed called the Québec Single-Family Building Stock Energy Model (QSFBSSEM). The QSFBSSEM accurately predicts the energy consumption of houses across the province of Québec, Canada. Approximately 1.9 million single-family dwellings are represented by the model. Space heating, water heating, lighting, appliances

and space cooling are predicted for detached and attached dwellings. Electricity, natural gas, heating oil and wood are the primary energy sources for residential buildings in the province and are also determined by the stock model. A total of 200,000 houses are modeled as a sample to represent the overall stock. The stock sample size is shown to be an important aspect for energy prediction of a building stock, as lesser-represented portions of the stock can result in significant deviation for the energy prediction if the sample size is too low. The QSFBSSEM can be used in a variety of applications and is applied to a case study for greenhouse gas emission (GHG) comparison between different heating system configurations for the building stock. The case study illustrates that the QSFBSSEM can predict the impact of large-scale conversion of non-electric heating systems on the peak electricity load of the building stock.



## TABLE OF CONTENTS

ACKNOWLEDGEMENTS .....	III
RÉSUMÉ.....	IV
ABSTRACT .....	VI
TABLE OF CONTENTS .....	VIII
LIST OF TABLES .....	XIV
LIST OF FIGURES.....	XVI
LIST OF ABBREVIATIONS .....	XX
LIST OF APPENDICES .....	XXII
CHAPTER 1    INTRODUCTION.....	1
CHAPTER 2    LITERATURE REVIEW .....	5
2.1    Energy modeling of a single-family dwelling.....	5
2.1.1    Energy balance in a SFD.....	5
2.1.2    Characteristics of single-family dwellings .....	9
2.1.3    Building energy models .....	13
2.2    Building stock energy modeling .....	16
2.2.1    Industry need .....	16
2.2.2    BSEM development techniques .....	17
2.2.3    Building archetypes.....	21
2.2.4    Examples of bottom-up white-box stock models based on physics simulation .....	25
2.2.5    Challenges in building stock energy modeling .....	27
2.3    Conclusion.....	29
CHAPTER 3    OBJECTIVES AND THESIS ORGANISATION .....	30
3.1    Thesis objectives .....	30

3.2	Main structure of the thesis .....	31
CHAPTER 4 ARTICLE 1: DEVELOPMENT OF A STOCHASTIC VIRTUAL SMART METER DATA SET FOR A RESIDENTIAL BUILDING STOCK - METHODOLOGY AND SAMPLE DATA .....		
		33
4.1	Abstract .....	33
4.2	Introduction .....	33
4.2.1	Smart meter data.....	34
4.2.2	Supervised machine learning classification of smart meter data .....	35
4.2.3	Electricity smart meter data available for classification studies .....	35
4.3	Objectives.....	37
4.4	Methodology .....	38
4.5	Virtual smart meter data generation .....	40
4.5.1	Development of a virtual smart meter framework .....	40
4.5.2	VSM framework details .....	41
4.5.3	Building geometry.....	42
4.5.4	Heating systems.....	43
4.5.5	Cooling systems .....	45
4.5.6	Lighting internal gains .....	46
4.5.7	Equipment internal gains.....	46
4.5.8	Domestic hot water.....	46
4.5.9	Air infiltration .....	46
4.6	Building parameters .....	47
4.6.1	Occupancy-driven internal loads.....	49
4.6.2	Uniform probability distributions.....	51

4.6.3	Probability mass functions .....	53
4.6.4	Fixed parameters .....	62
4.6.5	Window types.....	62
4.6.6	Climate files .....	65
4.7	VSM Data.....	65
4.7.1	Input data.....	65
4.7.2	VSM profile data.....	66
4.7.3	Load profiles .....	67
4.7.4	Annual totals for heating, cooling, lighting, equipment and domestic hot water electricity use.....	67
4.7.5	Overview of the data .....	67
4.8	Discussion .....	70
4.8.1	Developing and using a virtual smart meter data set .....	70
4.8.2	Classification modeling of electricity smart meter data.....	71
4.8.3	Automated load disaggregation algorithms.....	72
4.9	Conclusion.....	72
4.10	Acknowledgements .....	73
4.11	Data set.....	73
4.12	References .....	73
CHAPTER 5 ARTICLE 2: DISCRIMINANT ANALYSIS CLASSIFICATION OF RESIDENTIAL ELECTRICITY SMART METER DATA.....		78
5.1	Abstract .....	78
5.2	Introduction .....	78
5.3	Literature review .....	81

5.3.1	Machine learning in building applications .....	82
5.3.2	Smart meter analytics .....	83
5.3.3	Classification of smart meter data for building characterization .....	84
5.3.4	Linear discriminant analysis.....	85
5.3.5	Summary and paper organization.....	86
5.4	Data set description .....	87
5.4.1	Example data .....	91
5.5	Classification methodology.....	93
5.5.1	Predictive model development methodology .....	94
5.5.2	Classification accuracy.....	94
5.5.3	LDA predictive model development time .....	103
5.5.4	Impact of data set size on classification accuracy .....	104
5.5.5	Application of the developed predictive models on real smart meter data .....	106
5.6	General discussion.....	111
5.7	Conclusion.....	112
5.8	Glossary.....	114
5.9	References .....	116
	Appendix 5.1: Practical application of linear discriminant analysis .....	120
CHAPTER 6 ARTICLE 3: DEVELOPMENT OF A BOTTOM-UP WHITE-BOX BUILDING STOCK ENERGY MODEL FOR SINGLE-FAMILY DWELLINGS .....		127
6.1	Abstract .....	127
6.2	Introduction .....	127
6.2.1	BSEM approaches .....	128
6.2.2	Bottom-up white-box (BU/WB) building stock modeling.....	131

6.3	Objectives.....	134
6.4	Building stock characterisation .....	135
6.4.1	Dwelling types.....	135
6.4.2	Number of dwellings by region.....	136
6.4.3	Weather data.....	139
6.4.4	Building performance characteristics.....	141
6.4.5	System characterization.....	141
6.4.6	Occupancy .....	148
6.5	Model description.....	148
6.5.1	Building energy simulation .....	149
6.5.2	Stock energy consumption comparison.....	149
6.5.3	Building stock sample size.....	151
6.6	Case study .....	154
6.6.1	Greenhouse gas emission factors .....	155
6.6.2	Case study: 50% reduction in GHG emissions for space heating .....	157
6.7	Conclusion.....	159
6.8	References .....	161
	Appendix 6.1: Weather stations for each region of the studied building stock .....	165
	Appendix 6.2: Total and opaque cloud cover data correction.....	166
	Appendix 6.3: QSFBSSEM overview .....	168
CHAPTER 7	GENERAL DISCUSSION.....	172
7.1	Review of journal paper contributions .....	172
7.2	Other contributions.....	174
CHAPTER 8	CONCLUSION AND RECOMMENDATIONS.....	175

8.1 Recommendations .....	176
REFERENCES .....	178
APPENDICES .....	191

## LIST OF TABLES

Table 1.1 Residential building loads. See e.g. ASHRAE (2013).....	2
Table 2.1 Components of the indoor heat balance equation (ASHRAE 2013).....	7
Table 2.2 Building parameters .....	11
Table 2.3 Building stock modeling techniques: top-down versus bottom-up (Swan and Ugursal 2009).....	19
Table 2.4 Summary of recent bottom-up white-box building stock energy models (Q4 models)	26
Table 4.1 Summary of open-source residential smart meter data sets with relevant building information .....	36
Table 4.2 Model inputs and potential data sources .....	47
Table 4.3 Building parameters with uniform probability distributions.....	52
Table 4.4 Conditional probability distributions for roof thermal insulation based on wall insulation levels. Shaded values represent maximums .....	55
Table 4.5 Probability distributions for the stochastic parameters generated for each virtual building .....	57
Table 4.6 Window properties and clustering results .....	64
Table 4.7 Sample input set based on random number generation.....	65
Table 5.1. VSM data class category descriptions (adapted from Neale et al. 2020). PMF: probability mass function, UPD: uniform probability distribution (i.e. no prior knowledge for the building stock). Values in square brackets represent the median value for that category.....	88
Table 5.2. Classification accuracy results. $n_{cat}$ : number of category values for that class, RG: random guess, $RG_{PK}$ : random guess with prior knowledge. Results with a dark outline indicate the best result for that class .....	98
Table 5.3. RSM house characteristics. AC: air conditioning. Cat: class category according to Table 5.1 .....	107
Table 5.4. House data set known parameters and definitions for a correct and close prediction	108

Table 5.5. Classification accuracy results for the real smart meter data set. Best classification results have bold text and borders .....	110
Table 6.1 Examples of recent bottom-up white-box building stock energy models.....	131
Table 6.2 Distribution of occupied dwellings for Québec CMA areas. DF: dwelling fraction, SFD: single-family dwelling, Det: single-detached house, Row: row house, Semi: semi-detached house, OSA: other single-attached house (Statistics Canada 2016).....	138
Table 6.3 Heating system distribution for the Province of Québec for single detached and single attached homes (NRCan 2017) .....	142
Table 6.4 Fraction of homes by region based on primary heating energy (NRCan 2015) .....	143
Table 6.5 Heating system probability by region. Highlighting by data source from Table 6.4: electric, natural gas, wood or unknown.....	144
Table 6.6 Detailed heating system descriptions. O: heating oil, NG: natural gas, E: electric, W: wood. OAT: Outdoor air temperature .....	145
Table 6.7 Air conditioning prevalence in the studied building stock (NRCan 2015).....	147
Table 6.8 Domestic hot water system distribution by energy source (NRCan 2017).....	148
Table 6.9 Greenhouse gas emission factors for energy sources in the province of Québec .....	155
Table 6.10 Weather stations for each region of the studied building stock. CMA: census metropolitan area, CA: census agglomeration .....	165
Table 6.11 Example cloud cover data generated and filled for a 24 hour period. Night time values highlighted in grey are filled using linear interpolation between the two data points in bold text.....	167
Table 6.12 Overview of modeling choices for the building energy simulation of each house...	171
Table A.1 Review of building archetype research, adapted and updated from ( <i>Reinhart and Cerezo Davila 2016</i> ).....	192



## LIST OF FIGURES

Figure 1.1 Residential building types in the province of Québec, Canada, with an example dwelling identified in white.....	1
Figure 1.2 Average residential building energy consumption for Quebec ( <i>NRCan 2017</i> ) .....	3
Figure 2.1 Simplified depiction of the lumped capacitance model heat flows in a residential dwelling in heating mode. Red arrows show heat transfer to and from the indoor environment. HVAC: heating, ventilation and air conditioning .....	6
Figure 2.2 Distribution of house types in the province of Québec (Statistics Canada 2016) .....	9
Figure 2.3 Histograms for roof, foundation and wall thermal resistance values for a set of 27,000 houses (NRCan 2018) .....	10
Figure 2.4 Illustration of the top-down and bottom-up modeling strategies for a generic building stock .....	18
Figure 2.5 The quadrants for the energy dimension for the classification system proposed by Langevin et al. (2020). Additional dimensions based on population, environment and other factors also exist.....	20
Figure 4.1 Virtual smart meter data generation (Generator) and classification (Classifier) processes .....	39
Figure 4.2 Proposed framework including a manager and building energy model .....	42
Figure 4.3 Example of an occupant activity schedule over a two-day period for a 3-occupant home .....	51
Figure 4.4 Building parameter dependency network. R: thermal resistance of the building envelope, DHW: domestic hot water, HVAC: heating, ventilation and air conditioning.....	54
Figure 4.5 Window data by U-value and SHGC with resulting k-medoid clusters for single-, double- and triple-glazed windows. Size of EHD data points represents number of occurrences in the data set for a given window type.....	63

Figure 4.6 Box and whisker plot illustrating the variation in electricity consumption when compared with provincial averages.....	68
Figure 4.7 Interquartile ranges for the VSM data set and 30 houses in the province of Québec for (a) the first week of January, (b) the first week of July, and (c) daily energy use for a full year. ....	69
Figure 5.1. Generalized supervised machine learning predictive model development process.....	80
Figure 5.2. Annual electricity consumption of VSM data sorted by surface area category. Each point represents one house with distinct characteristics.....	91
Figure 5.3. Electricity consumption for houses with different characteristics for January (top) and July (bottom) hourly data .....	92
Figure 5.4. Confusion matrix for Scenario I (left) and Scenario IV for the Area class categories, labelled 1 through 5. TP: true positive, TN: true negative. Bolded values illustrate the correctly predicted cases.....	97
Figure 5.5. Classification accuracy per feature for the location, heated surface area, air infiltration and overall thermal resistance parameters. RG: random guess, RG <sub>PK</sub> : random guess based on prior knowledge.....	100
Figure 5.6. Comparison of January and July classification accuracy .....	102
Figure 5.7. Average model computation time based on the number of features.....	103
Figure 5.8. Area classification accuracy by building data set size for (a) monthly (12), (b) weekly (52), (c) daily (365) and (d) hourly (8760) features. b: number of buildings in data set, f: number of features, CA: classification accuracy.....	105
Figure 5.9. January and July electricity consumption for small and large houses. Class: house size, categories (2): small and large, features (2): January and July electricity consumption .....	121
Figure 5.10. Linear classification boundary with correct and incorrect predictions.....	124
Figure 5.11. Example of 5-fold cross-validation.....	126
Figure 6.1 Annual energy consumption (GJ/home) by end-use and energy source for detached and attached residential dwellings in the province of Québec, Canada (NRCan 2017) .....	128

Figure 6.2 Single-family dwelling (SFD) types in the province of Québec, Canada .....	136
Figure 6.3 Population distribution of the province of Québec, Canada, with CMA and CA regions superimposed. Adapted from Statistics Canada (2019). Approximate latitude lines are indicated for reference.....	137
Figure 6.4 Outdoor dry bulb temperature for 6 CMA locations across the province of Québec for CWECC weather data (left) and 2017 CWEEDs data (right).....	140
Figure 6.5 Model versus stock energy consumption for detached (DET), attached (ATT) and all single-family dwellings (SFD) .....	150
Figure 6.6 Box and whisker plot of the modeled single-family dwelling (SFD) energy consumption by end-use and energy source .....	151
Figure 6.7 NRMSD for energy consumption by end-use and energy source by fraction of the total building stock modeled .....	153
Figure 6.8 NRMSD for SFD energy consumption categories for cases representing 0.03% and 0.6% of the modeled building stock.....	154
Figure 6.9 Calculated CO <sub>2</sub> equivalent emission rates in December 2017 for electricity in the province of Québec and electricity usage by source for all sectors .....	156
Figure 6.10 Annual CO <sub>2</sub> equivalent emissions and total space heating energy consumption for the studied cases. Stock reference data from Natural Resources Canada is also provided for comparison (NRCAN 2017).....	158
Figure 6.11 Peak load percent difference for Scenarios 1 and 2 with respect to the base case scenario.....	159
Figure 6.12 General building stock energy modeling approach and model outputs.....	169
Figure 6.13 Building parameters generated for each house. Dependencies are illustrated with dashed arrows .....	170
Figure 7.1 Electricity consumption profile comparison tool.....	174

Figure A.1 Archetypes per number of buildings based on the literature .....	191
---	-----

## LIST OF ABBREVIATIONS

The following are some abbreviations commonly used throughout this work. Symbols are presented directly with each equation found within the thesis.

### **Abbreviations**

AC	Air conditioning
BSEM	Building stock energy model
CA	Classification accuracy
Cat.	Category
CM	Confusion matrix
CP	Correct predictions
DET	Detached house
DHW	Domestic hot water
EHD	Energuidé housing database
GHG	Greenhouse gas
HDD	Heating degree-days
HVAC	Heating, ventilation and air conditioning
LDA	Linear discriminant analysis
ML	Machine learning
MURBS	Multi-unit residential buildings
OSA	Other single-attached house
PK	Prior knowledge
PMF	Probability mass function
QSFBSSEM	Québec single-family building stock energy model
RG	Random guess
ROW	Row house
RSM	Real smart meter
SEMI	Semi-detached house
SFD	Single-family dwelling
SHEU	Survey of household energy use
SHGC	Solar heat gain coefficient

TP	Total predictions
UBEM	Urban building energy modeling
UPD	Uniform probability distribution
VSM	Virtual smart meter
WWR	Window-to-wall ratio

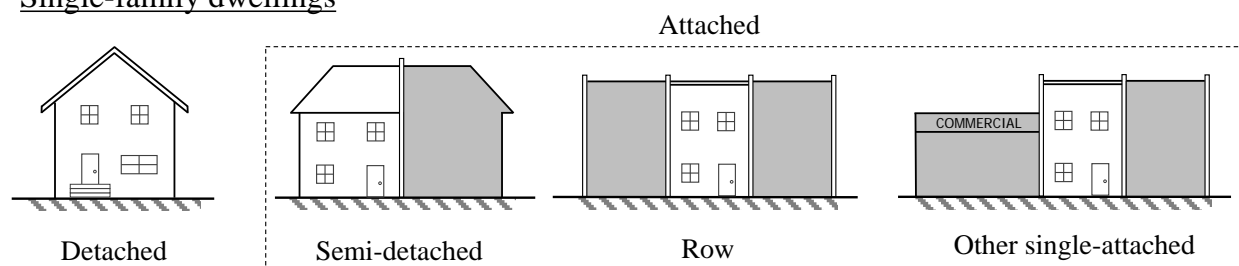
## **LIST OF APPENDICES**

Appendix A	Summary of building archetype research .....	191
------------	--	-----

## CHAPTER 1 INTRODUCTION

Residential, commercial and industrial buildings account for approximately 28% of all energy use in Canada and 22% of all greenhouse gas emissions (GHG) (NRCan 2017a). National and provincial objectives aiming to reduce energy consumption and GHG emissions target building energy use, among other measures, to reach long-term targets. For example, the province of Québec aims to reduce GHG emissions related to space heating by 50% by the year 2030 (Government of Québec 2020). Realistically meeting these objectives requires a deep understanding of building energy use at the building, regional, and provincial levels. The residential market represents more than half of the building energy consumption in Canada, and therefore understanding residential energy consumption is a logical starting point towards reducing energy use in the country.

### Single-family dwellings



### Multi-unit residential buildings

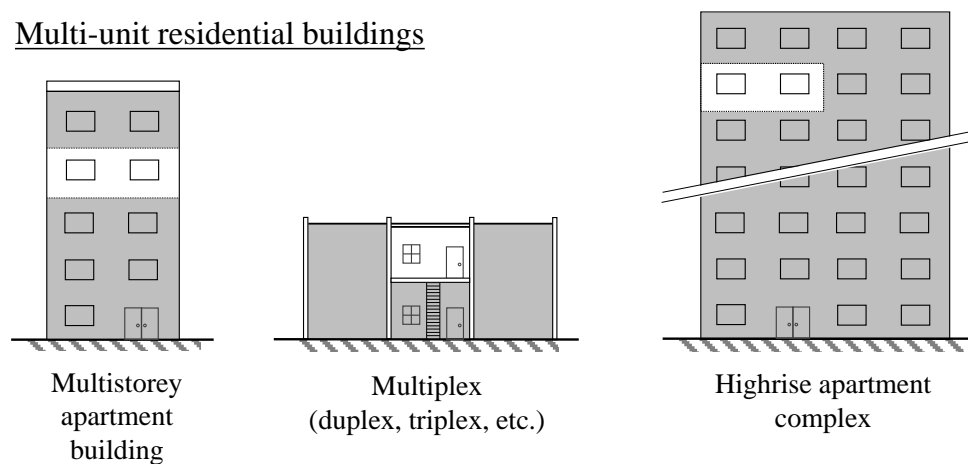


Figure 1.1 Residential building types in the province of Québec, Canada, with an example dwelling identified in white

The residential market is divided between single-family dwellings (SFD) and multi-unit residential buildings (MURBs). SFD are detached and attached buildings that house one family unit in a single



dwelling. Examples of SFD include detached (DET), semi-detached (SEMI), row (ROW) or other single-attached (OSA) houses (Figure 1.1). MURBS include buildings with multiple family units, such as apartment complexes or multiplexes, which house two or more families in separate apartments. Approximately 54% of all dwellings are of the SFD type, accounting for approximately 1.9 million homes in the province of Québec.

If the total energy consumption of a home is desired, the sum of all electricity and fuel requirements must be known. The various types of loads and some of the main factors driving the magnitude of those loads are presented in Table 1.1.

Table 1.1 Residential building loads. See e.g. ASHRAE (2013)

Load	Main drivers for the energy consumption	Description
<b>Heating</b>	<ul style="list-style-type: none"> <li>Outdoor temperature</li> <li>Thermostat setpoint</li> <li>Building envelope characteristics</li> <li>Heating system efficiency</li> <li>Internal gains</li> </ul>	Energy usage of the heating systems, e.g. electric baseboard heaters or a natural gas furnace. Typically electricity, natural gas, heating oil, wood, or other.
<b>Cooling</b>	<ul style="list-style-type: none"> <li>Solar gains</li> <li>Outdoor temperature</li> <li>Thermostat setpoint</li> <li>Building envelope characteristics</li> <li>Cooling system efficiency</li> <li>Internal gains</li> </ul>	Energy requirements for the cooling systems, e.g. an air conditioner. Typically electricity.
<b>Lighting</b>	<ul style="list-style-type: none"> <li>Natural lighting</li> <li>Occupancy</li> <li>Desired illuminance levels</li> <li>Type of lights</li> </ul>	Energy usage of the lighting, e.g. LED or incandescent light bulbs. Typically electricity.
<b>Equipment</b>	<ul style="list-style-type: none"> <li>Amount of equipment in the home</li> <li>Operation time of equipment</li> </ul>	Energy usage of: <ul style="list-style-type: none"> <li>Appliances;</li> <li>Electronics;</li> <li>Fans for ventilation;</li> <li>Other systems not covered by other categories, such as pools and portable spas.</li> </ul> Typically electricity with some natural gas and other energy sources.
<b>Domestic hot water</b>	<ul style="list-style-type: none"> <li>Hot water usage</li> <li>City water temperature</li> <li>Water heater properties (size, insulation, etc.)</li> </ul>	Energy usage of the water heater. Typically electricity, natural gas, or heating oil.

The average energy consumption by end-use for residential buildings in Quebec is illustrated in Figure 1.2. Heating energy consumption represents roughly two-thirds of energy use for a typical building. Occupancy-driven loads such as lighting, equipment and domestic hot water (DHW) represent the other one-third.

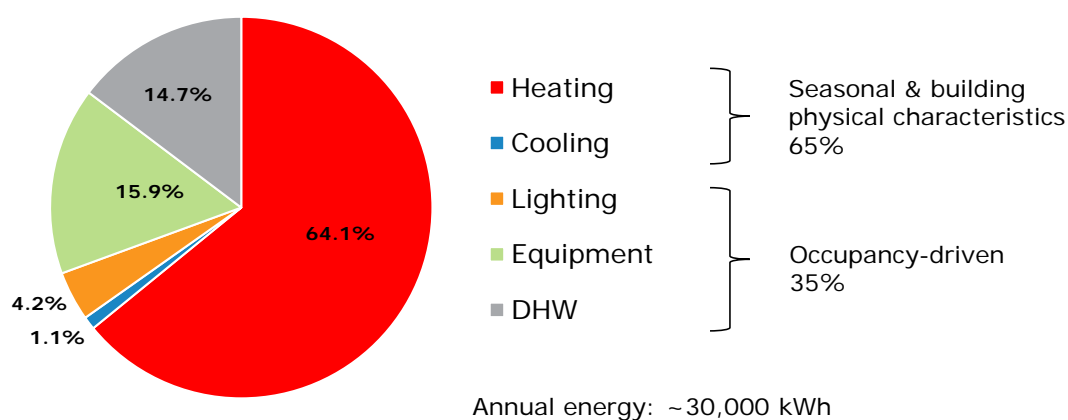


Figure 1.2 Average residential building energy consumption for Quebec (*NRCan 2017*)

The energy distribution in Figure 1.2 provides an average portrait of residential dwelling energy consumption. This average representation does not illustrate the variability from one house to another that can result from the heating system efficiency, the number of occupants, the energy source used for heating, the thermal performance characteristics of the house, the behaviour of the occupants, etc. These factors all influence the overall energy use of a home, and will determine the impact of an energy conservation measure on the regional or provincial energy consumption. A *building stock energy model* is required to accurately evaluate a widely implemented efficiency measure or other feature on the residential building stock.

The percentage distribution illustrated in Figure 1.2 depends on the various parameters described in Table 1.1, which in turn depend on the energy source used for each end-use in the home. The result is a very large set of possible input parameters that must be described, or *characterized*, in order to develop an accurate building stock energy model.

There are very few building stock energy models currently available, with even less applicable to the Canadian building stock. A workshop held in 2017 on building stock modeling identified an urgent need for additional tools for large scale building stock energy analysis and data (Ugursal

2017). This study provides new solutions to address the limited availability of accurate, reliable stock energy prediction tools that are required for proper evaluation of energy conservation measures, GHG emission studies and a variety of other urgent needs for government and industry stakeholders. This thesis strives to solve some of the inherent limitations often associated with the development of building stock energy models, such as the lack of data for describing the variety of building parameters required for an accurate model, or the restrained stock model size due to limited computational resources. In addition, a new residential stock energy model is provided for single-family dwellings in the province of Québec, Canada.

## **CHAPTER 2      LITERATURE REVIEW**

Building stock energy modeling (BSEM) is a complex endeavour requiring a deep understanding of the physical processes involved at the building-level. A model of a stock is a sum of individual parts, and consequently the overall precision will depend on the accuracy of each part. In order to discuss the advantages and limitations associated with BSEMs, the process of modeling the energy use of a single-family dwelling is first described, from the internal loads to the physical parameters that make up a building. The link between the physical processes occurring in a building and the actual characteristics of a building is described. The overall energy consumption outputs of a house are presented and linked to the internal loads, where applicable.

Once energy modeling at the building-level is fully outlined, the process of stock modeling is presented. The advantages and limitations of various BSEM methods are described in detail, with certain methods elaborated in more detail. Finally, a summary is provided outlining the key points leading to the need for the present study and the approach that is adopted.

### **2.1 Energy modeling of a single-family dwelling**

A house provides shelter for one or more occupants from undesirable outdoor elements, such as rain, pests and harsh outdoor conditions. Since the focus of this study is on energy use, the insulative properties of a dwelling's exterior surfaces are studied in particular. A house and the various systems therein must be able to keep an occupant at a thermally comfortable temperature. The processes affecting the indoor environment in a typical home are described for a generic single-family dwelling (SFD). The link between the building's properties and those internal and external heat flows is established. Various building energy modeling tools are described, with an emphasis on the desired outputs and the general process of modeling a SFD.

#### **2.1.1 Energy balance in a SFD**

The indoor environment of a house is separated from the outdoor conditions by a building envelope. The conditions in the house vary over time as various loads act on the indoor air. The result is a transient set of processes that determine the incremental change in temperature over a certain time step. While a variety of methods exist to express the system of equations required to determine the

transient temperature change in a building, the lumped capacitance method is a simple technique that is applicable for simple cases such as single-family dwellings.

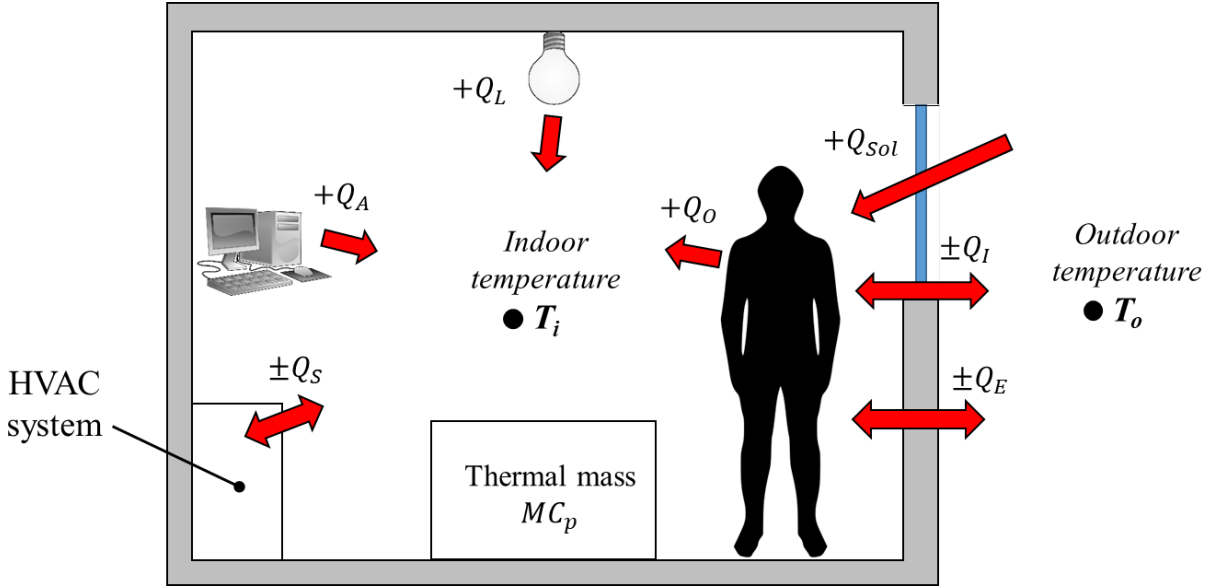


Figure 2.1 Simplified depiction of the lumped capacitance model heat flows in a residential dwelling in heating mode. Red arrows show heat transfer to and from the indoor environment.

HVAC: heating, ventilation and air conditioning

A simplified version of the lumped capacitance model (LCM) is described in Equation (2.1) for the case illustrated in Figure 2.1. Modern building simulation software programs use more sophisticated methods (see e.g. the TRNSYS mathematical library (TRNSYS 2017)), but the LCM is based on the same foundational elements and serves the purpose of illustrating the heat balance in an indoor space. The heat gains and losses affecting the indoor environment are equated to the thermal mass component, which represents the storage of heat in the contents of the home and allows the determination of the transient temperature change. Furniture (couches, tables, chairs, appliances, etc.), finishing materials (carpets, wood floors, etc.), construction materials (brick, wood structural components, insulation, etc.) and other contents all contribute to the thermal mass in a home.

$$Q_L + Q_O + Q_A + Q_{Sol} \pm Q_S \pm Q_E \pm Q_I = MC_p \frac{\partial T}{\partial t} \quad (2.1)$$

The principal components of Equation (2.1) are described in more detail in Table 2.1. The magnitude of the heat gains and/or losses depend on a variety of physical parameters of the building and the surrounding environment. The equations for each component are not presented here for brevity but can be found in many industry and academic references. In the case of this study, the ASHRAE Fundamentals (ASHRAE 2013) reference is used for specific details related to the physical processes described in Table 2.1.

Table 2.1 Components of the indoor heat balance equation (ASHRAE 2013)

Var.	Description	Units
$Q_L$	The lighting capacity in the zone, which emits heat to the indoor air via convection and long-wave radiation.	W
$Q_O$	Heat emitted by occupants, transferred based on convective and radiative fractions. Additionally there are sensible and latent portions of the heat, as occupants generate moisture in the indoor environment. The amount of total heat depends on activity level, gender, age and other factors.	W
$Q_A$	The heat gain to the indoor environment by electronic and other appliances, including domestic hot water systems, appliances such as refrigerators, dishwashers, stoves, and electronic devices such as computers, laptops, etc. Heat gain depends on the type of appliance, which have specific convective and radiative fractions. Occupancy greatly influences the frequency and duration of use for appliances.	W
$Q_S$	The heating or cooling in the space, which at any given moment typically do not operate simultaneously in residential buildings. These systems are sized to be able to maintain an equilibrium in the indoor environment such that the indoor temperature can be maintained at a comfortable level for occupants. In heating mode $Q_S$ is positive, i.e. heat is provided to the indoor environment. In cooling mode $Q_S$ is negative.	W
$Q_{Sol}$	The solar gains in the house, which are a function of the intensity of incident rays, the angle of the sun, the properties of the windows, such as the solar heat gain coefficient, and the surfaces in the house, such as the emissivity of the materials.	W
$Q_E$	The heat gains or losses via the building envelope, including the walls, roof, windows, etc. The amount of heat lost or gained to the indoor environment is a function of the properties of the building envelope, such	W

Var.	Description	Units
	as the thermal conductivity and thickness of the materials. In heating mode $Q_E$ is negative, i.e. heat lost to the outdoor environment, and in cooling mode $Q_E$ is positive. The total gains or losses are a function of the temperature difference between indoors and outdoors and the thermal resistance of each component of the building envelope.	
$Q_I$	The heat gains or losses due to outdoor air infiltrating to or from the indoor environment. Total infiltration is a function of many parameters, such as the total leakage area, outdoor temperature, the pressure differential across the building envelope, the wind speed and direction, etc. Specific air infiltration models are often used to represent air infiltration, such as the Sherman-Grimsrud method (Sherman and Grimsrud 1980). In heating mode $Q_I$ is negative and in cooling mode it is positive.	W
$MC_p$	The lumped capacitance for the indoor environment, which is a representation of the sum of the mass and heat capacitance of every single object in the space. The lumped capacitance acts to mitigate rapid temperature fluctuations in a space as heat is stored and released in objects.	$\text{kJ} \cdot ^\circ\text{C}^{-1}$
$\frac{\partial T}{\partial t}$	The rate of change of temperature in the indoor environment over time. In practical scenarios, the component $Q_S$ is controlled with the use of a thermostat to minimize the change in temperature in the indoor space, which maintains a comfortable environment for occupants.	$^\circ\text{C} \cdot \text{s}^{-1}$

The heat gains and losses illustrated in Figure 2.1 and described in more detail in Table 2.1 provide an initial understanding of the types of physical characteristics required to determine the rate of change of the temperature in an indoor environment of a dwelling. For example, the air infiltration depends on the leakage area of a dwelling, the wind speed, outdoor air temperature, and a variety of other factors. There is naturally a direct link between the physical properties and the resulting energy consumption of a building.

Representing the variety of building characteristics influencing the overall energy consumption for a building stock requires an understanding of how they vary from one house to another. The information available varies widely depending on the type of building (residential, commercial, industrial, etc.), the type of stock (urban, regional, national, etc.), and government data collection

programs that are available. These factors are addressed for the studied building stock in the following section.

### 2.1.2 Characteristics of single-family dwellings

As described in the introduction of the thesis, single-family dwellings (SFDs) represent a significant contribution towards the national energy consumption in Canada. They are targeted as a first effort towards developing a full residential building stock model for the province of Québec, and eventually, Canada. Single-family dwellings in the province of Québec fall into two broad categories: detached and attached dwellings, depicted in Figure 1.1. Detached (DET) dwellings are standalone buildings typically one- or two-storeys, most often with finished basements. The size of detached homes across the province varies significantly, though on average they are larger than attached houses. Attached homes are divided into three subcategories: 1) row houses (ROW), 2) semi-detached (SEMI) houses, and 3) other single-attached (OSA). Semi-detached houses share an external boundary with another dwelling, but usually have three sides exposed to the outdoor environment. Similarly, row houses have two shared surfaces with adjacent dwellings, as do other single-attached dwellings, with the difference being one or both of the adjacent buildings is non-residential in the case of OSA houses. The proportion of dwelling types in Québec is illustrated in Figure 2.2.

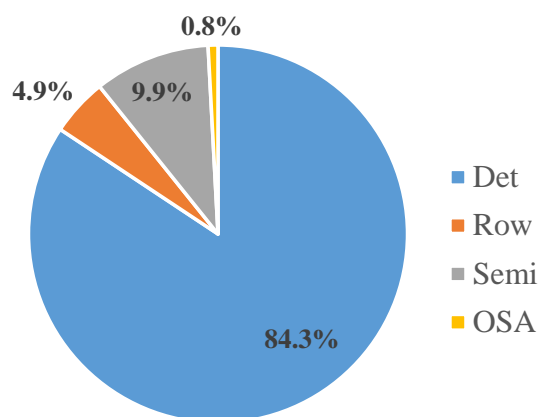


Figure 2.2 Distribution of house types in the province of Québec (Statistics Canada 2016)



Detached homes are by far the most common type of SFD in the province of Québec. Data on the dwelling type distribution and occupancy is available for every census division in the province from Statistics Canada (2016).

House construction varies based on construction year, house type and retrofit level, resulting in a variety of wall, roof and foundation configurations. While the materials used are relevant for thermal mass considerations, overall insulation level is an important factor to consider. The insulative properties of the components of a particular building envelope assembly, such as an exterior wall, are typically combined and expressed in terms of *thermal resistance* or *RSI*, as in R-value in the International System of Units (SI). Thermal resistance increases proportionally with the thickness of the insulation in a building envelope assembly.

As an example of building properties available in the literature, the Energuide retrofit program contains a database of house parameters collected over almost 30 years of house retrofits. Over 700,000 homes across Canada are represented, including 26,000 in the province of Québec. The thermal resistance values for wall, roof and foundation assemblies of the Energuide homes in Québec are illustrated in Figure 2.3.

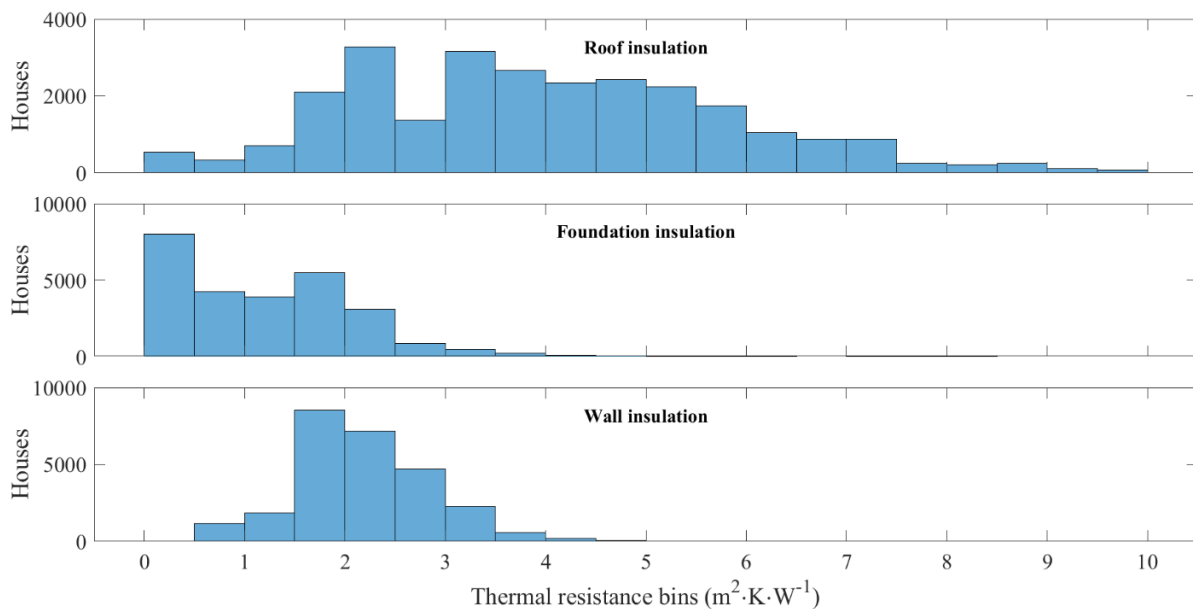


Figure 2.3 Histograms for roof, foundation and wall thermal resistance values for a set of 27,000 houses (NRCan 2018)

The insulation levels vary significantly across each category in Figure 2.3 and provide an example of the variable nature of building properties in the studied building stock. Additional properties, such as the leakage area, heated surface area and window type, are also available in the Energuide database.

Based on the heat balance for the indoor environment presented in Section 2.1.1, a list of building parameters is provided in Table 2.2. While not an exhaustive list, the parameters provide an overview of the main input requirements required to simulate the indoor conditions in a dwelling. The identified dependencies of each parameter are provided, in addition to a source of data for the parameter, if applicable.

Table 2.2 Building parameters

Property	Dependencies	Source for data	Description
<b>Location and geometry</b>			
<b>1. Region</b>	N/A	Census (Statistics Canada 2016), SHEU (NRCan 2011), SHEU (NRCan 2015)	Region in Québec, determines the local weather conditions
<b>2. Building type</b>	Region	Census (Statistics Canada 2016), SHEU (NRCan 2011), SHEU (NRCan 2015)	Single detached, semi-detached, row, etc.
<b>3. Surface area</b>	Building type	SHEU (NRCan 2011), SHEU (NRCan 2015)	Heated surface area of a home including finished basements.
<b>4. Shape</b>			Aspect ratio, e.g. 0.8, 1.0, 1.3
<b>5. Adjacency</b>	Building type		Type and number of shared external boundaries with other buildings.
<b>6. Rotation</b>			0° to +270°
<b>7. # Floors</b>	Building type	SHEU (NRCan 2011)	One or two floors
<b>8. Window to wall ratio (WWR)</b>			e.g. 0.1, 0.2, etc.

Property	Dependencies	Source for data	Description
<b>Thermal performance</b>			
<b>9. Wall construction</b>		EHD (NRCan 2018)	Level of insulation
<b>10. Roof construction</b>	Wall construction	EHD (NRCan 2018)	Level of insulation
<b>11. Foundation construction</b>	Wall construction	EHD (NRCan 2018)	Level of insulation
<b>12. Infiltration</b>	Building age Location Weather HVAC	EHD (NRCan 2018)	Rate of infiltration
<b>13. Window type</b>		SHEU (NRCan 2011)	Number of glazings, solar heat gain coefficient, U-value
<b>Systems</b>			
<b>14. Heating system</b>	Location Building type	SHEU (NRCan 2015), NEUD (NRCan 2017b)	Heating system details (type, energy source, efficiency, etc.)
<b>15. Cooling system</b>	Heating system	SHEU (NRCan 2015), NEUD (NRCan 2017b)	Cooling system details (type, efficiency, etc.)
<b>16. DHW system</b>	Heating system	SHEU (NRCan 2015), NEUD (NRCan 2017b)	DHW system details (type, energy source, efficiency).
<b>17. Pool</b>	Building type	Hydro-Québec (Hydro-Québec 2019a)	Presence and type of pool
<b>18. Spa</b>		Hydro-Québec (Hydro-Québec 2019b)	Presence and type of portable electric spa
<b>Occupancy and internal loads</b>			
<b>19. Appliances</b>	Building type Occupants	CREST demand model (McKenna and Thomson 2016)	Number and type of appliances
<b>20. Lighting</b>	Building type	CREST demand model (McKenna and Thomson 2016)	Type of lighting, density
<b>21. Appliance usage</b>	Appliances Occupancy	CREST demand model (McKenna and Thomson 2016)	Equipment electricity consumption

Property	Dependencies	Source for data	Description
<b>22. Lighting usage</b>	Lighting Occupancy	CREST demand model (McKenna and Thomson 2016)	Lighting electricity consumption
<b>23. Occupants</b>	Building type Location	CREST demand model (McKenna and Thomson 2016), Census (Statistics Canada 2016), SHEU (NRCan 2011)	Number of occupants
<b>24. Occupancy</b>	Occupants Location Building type Time Day	CREST demand model (McKenna and Thomson 2016)	Occupant schedule and activity schedule.
<b>25. DHW draws</b>	Occupants Occupancy Time Day	CREST demand model (McKenna and Thomson 2016)	Hot water draw profile

While other details are necessary to describe a single-family dwelling's energy consumption, the parameters presented in Table 2.2 provide an overview of the most relevant and influential characteristics on the overall energy use of a house. Once known, the building parameters can be used to build up a building energy model of a house, for determining heating and cooling loads and overall building energy consumption.

### 2.1.3 Building energy models

The process of combining all of the building loads into a system of equations and solving for the energy consumption of the various components is not a new concept. Many tools exist that calculate the loads and energy consumption of a building. A review of building simulation programs in 2008 put the number of tools available in the hundreds (Crawley et al. 2008). There are professional certification programs available specifically for building energy modelers, such as the American Society of Heating, Refrigeration and Air Conditioning Engineers (ASHRAE) Building Energy Modeling Professional (BEMP) certification (ASHRAE). This illustrates that building energy modeling is accepted as an industry practice and not just as a research tool.

The difficult part of building simulation is not using a tool to produce energy consumption results, it is trying to reproduce the energy consumption of a real building exactly. There is a relatively

famous quote by statistician George E. P. Box that reads “all models are wrong, but some are useful” (Box 1976). An example of this is attempting to model the exact behaviour of a family in a house. Occupants are very stochastic in nature and behaviour can vary widely from one family to another based on a variety of demographic factors. A model will contain assumptions and will therefore never correspond perfectly to reality. However, a properly prepared model can nevertheless be a useful tool during the design and operation stages of a building.

### **2.1.3.1 Applications for building energy modeling**

Higgins (2012) identified four main purposes for building energy simulation in industry:

- Code compliance – compares a building design to a reference building to determine code compliance
- Design performance – energy performance simulations performed for building labeling programs
- Measurement and verification – verification of energy savings for building upgrades during the operation phase
- Building asset ratings – commercial building modeling for building asset ratings.

For research purposes, building energy modeling can cover a much wider spectrum of possible applications. Researchers often create their own tools for specific building energy modeling applications. Some examples of tools used in industry and/or research environments include:

- TRNSYS, a modular transient system modeling tool with a large library of building and HVAC components that can be combined together to create complex models (Klein et al. 2017)
- EnergyPlus, a building energy modeling program allowing detailed design load and energy consumption calculations for all types of buildings (Crawley et al. 2001)
- DOE-2.x, a free building energy modeling and cost analysis program, and the more advanced version eQuest developed using DOE-2 as an engine (JJH & LBNL 2021)
- ESP-r, a building performance simulation research tool used primarily in Europe and Canada for building energy prediction (“ESP-r: Building Performance Simulation Tool.”)
- The CREST thermal response model, which provides high-resolution stochastic occupancy loads and building energy use for residential buildings (McKenna and Thomson 2016)

The list above is by no means a comprehensive review of building energy modeling programs. Crawley et al. (2008) provide an overview of the most popular energy models used in academia and industry. For most research applications, the choice will be based on personal preference and

certain specific model capabilities. For a project such as this one, where residential buildings are the targeted building types, many programs could be considered adequate. The TRNSYS building energy simulation software is used throughout this study as it has the required capabilities to complete the project.

### 2.1.3.2 Model outputs

Once a building energy model is selected, the purpose of the simulation must be established. In the case of this study, the energy consumption of the various systems in single-family dwellings is targeted. National energy use data is often presented in the following five categories, such as the data from Figure 1.2 (NRCan 2017b):

1. **Space heating.** The total energy use for heating, which in the studied building stock uses primarily electricity, natural gas, heating oil and/or wood. Hybrid systems using more than one energy source are possible.
2. **Space cooling.** The total energy use for cooling. Unlike for heating, many houses do not have cooling systems in the studied building stock, and therefore this value can often be equal to zero. In the studied building stock, cooling systems are 100% electricity-based.
3. **Appliances.** The total energy use of systems in and around the dwelling, excluding domestic hot water systems, lighting, heating and cooling systems. Primarily electricity, though some non-electric appliances are possible.
4. **Lighting.** The total energy use of lighting systems in a dwelling. Lighting is 100% electricity-based.
5. **Water heating.** The total energy use of domestic hot water systems in a dwelling. Electric, natural gas and heating oil systems are possible in the studied building stock.

By separating the electric and non-electric energy sources the total peak electricity use of a house can be determined by summing the various electric systems. The total end-use energy can also be determined based on each energy source: electricity, natural gas, heating oil and wood. To be comparable with electricity smart meter data, which is recorded at 15-minute intervals for the

studied building stock (Hydro-Québec 2012), model energy data can be output at the same frequency.

## 2.2 Building stock energy modeling

A building stock is defined here as a group of buildings sharing some common root characteristic. This could be the same building type (i.e. residential), the same geographic location (i.e. a community, city or province) or a combination of those traits. A building stock model does not need to represent all of the buildings in the same geographic location. For example, the energy consumption of all residential buildings in a province could be required for a demand-side management effort by a local energy distributor targeting the residential sector.

A building stock energy model (BSEM) is a model that represents the energy consumption of a building stock. It should be noted that the term urban building energy modeling (UBEM) is sometimes used synonymously with BSEM in the literature (Reinhart and Cerezo Davila 2016). However, the term *urban* specifically refers to groups of buildings on the order of “thousands”, typically targeting a market at a city scale. The buildings for UBEM are also not necessarily restricted to residential single-family dwellings as multifamily dwellings or commercial buildings could be included. For this study, the BSEM terminology is adopted.

In order to present building stock energy models and associated works in the literature, the industry need for such tools is first discussed. The basic techniques for developing a BSEM are then presented. Building archetypes as a stock modeling approach are presented, with specific emphasis on the segmentation and characterization processes used to develop archetypes. Finally, some example building stock energy models are provided for reference to existing works in the field, followed by the challenges associated with BSEM development.

### 2.2.1 Industry need

Many circumstances require the energy prediction of large collections of buildings. Some applications for BSEM include, but are not limited to, the following items (Ugursal 2017):

1. National energy use projections
2. National efficiency regulation planning

3. Codes and standards development
4. Greenhouse gas emissions calculations
5. Renewable energy system design
6. Energy transportation and distribution network planning
7. Energy demand management
8. Demand-side management studies
9. District heating system design
10. Urban building energy modeling
11. City planning
12. Etc.

In some cases, the applications above may only require high-level energy consumption details for the buildings in the considered stock. For example, for national energy use projections for residential buildings, annual energy use would likely be sufficient. For a community expansion project, a local utility would be interested in total energy use and peak usage (maximum power use for the community for a given period) for different moments of the year.

Despite the high number of potential applications for BSEM and the high demand from industry there exists few appropriate models to achieve those goals in Canada (Ugursal 2017). Simulating every building individually in a building stock is a time-intensive process. An annual simulation for a single residential building can take hours to develop and up to several minutes to run. Whether it is for national residential energy prediction (over 13 million households in Canada (Statistics Canada 2016)) or for a community of 1,000 homes, it is clear that a simplified method of modeling building stock energy use is likely required.

### **2.2.2 BSEM development techniques**

Two principal techniques for representing building stock energy use exist: top-down and bottom-up models. Top-down models start at the figurative “top level” of the building stock, represented by the average energy consumption of all buildings of that type. The buildings could then be categorized in various ways (for example, in the decision-tree manner illustrated in Figure 2.4) until the individual building level is reached. Bottom-up models represent the individual buildings first in order to characterize the energy use of a larger group of buildings (for example, a community). The notation of each segment of the tree denotes the subgroup of buildings, i.e.



$n_{c1}|(n_b, n_{bt2})$  indicates the homes in climate category  $c1$  given market  $b$  and given building type  $bt2$ .

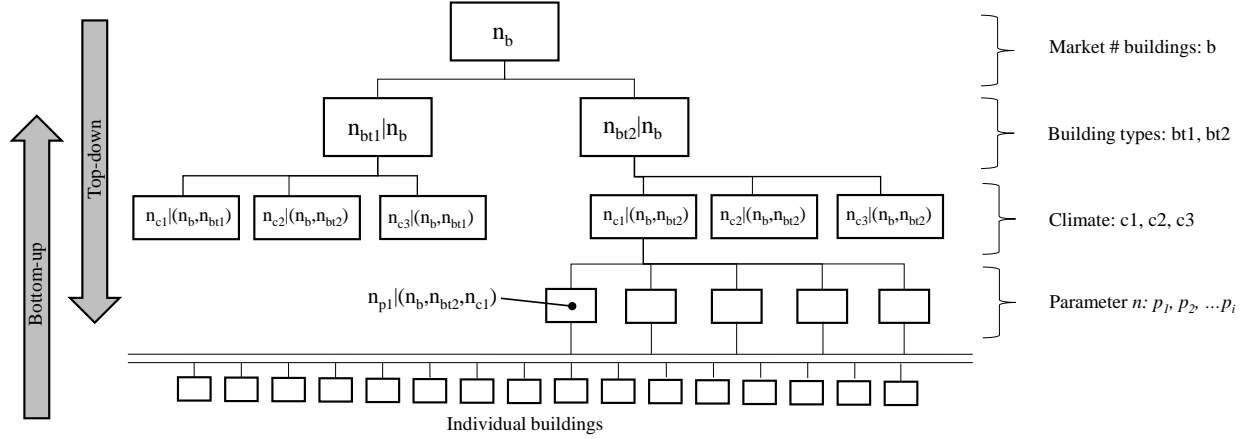


Figure 2.4 Illustration of the top-down and bottom-up modeling strategies for a generic building stock

Swan and Ugursal (2009) performed a comprehensive review of residential building energy modeling techniques. The bottom-up category of stock modeling is further divided into two subcategories: statistical models, which are based on the real energy use of buildings, and engineering models, which are based on physical processes in buildings. The three bottom-up modeling categories identified by Swan and Ugursal are summarized in Table 2.3. While top-down models are presented, they are not the focus of this study and are not discussed in further detail. This is because top-down models are not apt to evaluate technological changes to the building stock, which is one of the interests of this study.

Table 2.3 Building stock modeling techniques: top-down versus bottom-up (Swan and Ugursal 2009)

<b>Modeling technique</b>	<b>Description</b>	<b>Advantages</b>	<b>Limitations</b>
<b>Top-down</b>	<p>Future energy use is predicted based on historical energy use and building parameters. Addresses the whole building stock together. Further categorizations:</p> <ul style="list-style-type: none"> <li>• Econometric</li> <li>• Technological</li> </ul>	Adequate for predicting the future impact of small changes to the building stock.	Inaccurate for large change prediction, such as significant changes to building codes or large changes in the housing stock.
<b>Bottom-up (statistical)</b>	<p>Based on historical energy use which is tied to building characteristics. Further categorizations:</p> <ul style="list-style-type: none"> <li>• Regression</li> <li>• Conditional demand analysis</li> <li>• Neural networks</li> </ul>	Results are based on real building energy consumption, and therefore more realistic.	Results depend on the accuracy and quantity of the data and may not represent the full extent of the building stock.
<b>Bottom-up (engineering)</b>	<p>Explicitly accounts for building energy consumption based on equipment and modeling of physical processes in the building. Further categorizations:</p> <ul style="list-style-type: none"> <li>• Population distribution</li> <li>• Archetypes</li> <li>• Sample</li> </ul>	<p>Can account for all systems within a building, provided they can be defined in a model. More easily accounts for changes within those systems to test differences, such as studying energy efficiency upgrades for heating systems.</p>	Requires many assumptions and is limited by the accuracy of models.

The bottom-up methods identified by Swan and Ugursal (2009) and presented in Table 2.3 are further categorized into specific techniques. Statistical techniques rely primarily on fitting correlations and regression curves to building energy data, or by creating neural networks that seek to explain the relationship between building parameters and energy consumption. Engineering techniques rely on stock simplification techniques aided by building energy simulation and other physical models. The key difference between statistical and engineering techniques is the fact that the former relies on actual building data, which can add accuracy to the outcome, but does not as easily adapt to changes in technology that are not represented in the data set. In reality, statistics-based techniques will always be constrained to the bounds of the parameters contained in the data set. Engineering-based cases are constrained by the inputs of the modeler but have more flexibility.

Langevin et al. (2020) more recently elaborated further on the definitions outlined in previous works, seeking to establish a universal language for describing building stock models. Their work is based on a collective effort of the International Energy Agency (IEA) Annex 70 project on Building Energy Epidemiology (IEA 2021). Langevin et al. adopt a different terminology to express engineering- versus statistics-based models. Cases based on historical data are *black-box* models, while those that rely upon detailed physical characterisation are *white-box* models. The top-down and bottom-up terminology is also retained by Langevin et al, resulting in four principal so-called *quadrants* of building stock energy model, with a fifth representing a mix of techniques from two or more of the quadrants, i.e. *hybrid* models (Figure 2.5).

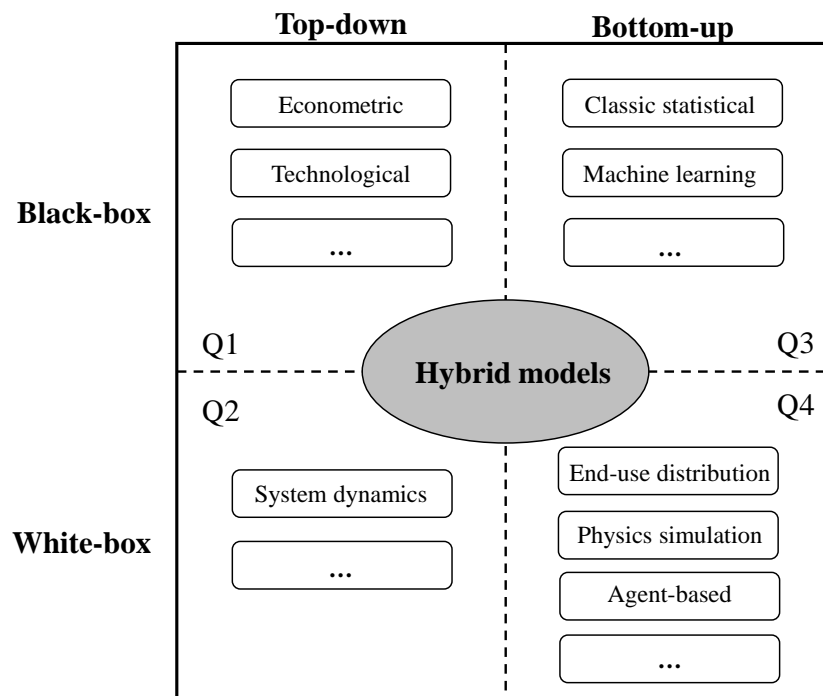


Figure 2.5 The quadrants for the energy dimension for the classification system proposed by Langevin et al. (2020). Additional dimensions based on population, environment and other factors also exist

While the new approach by Langevin et al. presents numerous differences over the category descriptions by Swan and Ugursal (2009), the most relevant to this study is the categorization of *Physics Simulation* bottom-up white-box energy models. Physics Simulation encompasses building

archetypes and geo-spatial models that rely on an underlying building simulation to provide the energy consumption of the stock components. Building archetypes are a common approach taken to simplify stock energy prediction, while geo-spatial models represent buildings in the physical environment where they are located using techniques such as geographic information systems (GIS). Due to the prevalence of building archetypes and the relevance of the techniques used to develop them, archetypes are presented in further detail in Section 2.2.3.

One of the contributions of the work by Langevin et al. is a categorization system for new building stock energy models. Newly developed models are proposed to be presented based on the following details (J. Langevin et al. 2020):

1. General purpose of the model.
2. Quadrant classification, as described in Figure 2.5.
3. Modeling technique, such as statistical, machine learning, physics simulation, etc.
4. Other quadrant classification layers, such as population or environment, if applicable, and the treatment of additional dimensions.
5. Availability of the model and associated details (i.e. publications) related to the development and use of the model.

Certain member countries of IEA Annex 70 have now presented their proper building stock models in terms of the classification system identified above, which are provided in the work by Langevin et al. (2020).

As a final note, a distinction is made between *static* and *dynamic* models. Static models represent a building stock in the present state (i.e. the status quo), and are used for energy assessment and comparative analysis using the current stock composition. Dynamic models are used for future prediction of energy consumption of a building stock (J. Langevin et al. 2020).

### **2.2.3 Building archetypes**

Building archetypes are templates used to simplify the energy consumption prediction for building stocks. In Europe the term *building typology* is used synonymously with building archetype (Loga,

Stein, and Diefenbach 2016). Building archetypes are intended to contain characteristics that allow an energy modeller to represent the buildings within a stock without representing each individual case. A review of previous works related to archetype development is provided in Appendix A for reference.

To represent buildings as archetypes, a building stock must go through a *segmentation* and *characterization* process (Sokol et al. 2016). While the term *classification* has been coined for building archetypes by some authors, in this study the term *segmentation* will be used to avoid confusion with the machine learning process also known as *classification*.

### 2.2.3.1 Segmentation

Segmentation is the process of categorizing buildings based on their physical parameters. For example, some archetypes segment stocks based on building age, HVAC systems, building shape, etc. (Reinhart and Cerezo Davila 2016). The process of dividing the building stock depends on the parameter and the information available. In some cases detailed geographic information system (GIS) data is available for an urban building stock, and therefore building shape does not need to be included in the archetypes. In addition, in urban stocks and small countries, all buildings can be subject to the same climate, and therefore there is no reason to segment based on climate. In countries such as Canada, the wide diversity in climate requires archetypes to include some sort of regional segmentation. One could imagine that two identical buildings with identical occupants, one in Vancouver and one in Whitehorse, would have very different annual energy consumption profiles. They could therefore not be accurately represented by a single archetype.

Parameters that could be used to segment a building stock include (see e.g. (Reinhart and Cerezo Davila 2016)):

- Climate
- Shape (surface area, height, aspect ratio)
- Construction characteristics (wall assembly, windows, window-to-wall ratio)
- Construction year (building age)
- Application (residential, commercial, etc.)
- Building type within an application (detached house, semi-detached house, row house)
- Systems (different mechanical system components and efficiencies)

### 2.2.3.2 Characterization

Once a building stock is segmented based on certain characteristics, those categories must be assigned ranges of values. Continuing the example of climate from the *segmentation* process, Canada could be divided into regions based on heating degree-day brackets. Parameters are characterized either with information from the literature, such as the Canadian Survey of Household Energy Use, or taken from actual building data, which represents only a sample of the larger building stock (Sokol et al. 2016).

The maximum number of archetypes possible for a segmented and characterized building stock is simply the multiplication of the number of each characterization, as expressed by Equation (2.2).

$$Arch_{Max} = \prod_{i=1}^s k_i \quad (2.2)$$

where:  $s$  is the number of segmentation parameters

$k$  is the number of characterizations for each segment.

Note that Equation (2.2) presents the maximum number of potential archetypes. Some combinations of parameters may not realistically represent any buildings in the building stock, which can be eliminated using a variety of graphical and numerical techniques (Famuyibo, Duffy, and Strachan 2012). In more practical terms, a variety of techniques have been used to characterize building stocks, which include but are not limited to:

- Literature review, such as analysis of building codes (e.g. Sokol et al. (2016))
- Graphical methods, such as frequency histograms and clustering of key building parameters (e.g. Famuyibo et al. (2012))
- Statistical techniques, such as calibration using real energy consumption (e.g. Sokol et al. (2016))
- Sampling of real building characteristics (e.g. Ballarini et al. (2014))

As a general rule, those developing archetypes for a building stock make use of the information that is available to best characterize the building stock. Works such as those by Sokol et al. introduce additional techniques (such as Bayesian calibration) to further improve the accuracy of archetypes, because simply using the available deterministic building properties found in the literature was deemed insufficiently accurate for their building stock (Sokol et al. 2016).

### **2.2.3.3 Energy prediction accuracy of building archetypes**

The accuracy of a building stock model is typically validated against measured data. Studies have either validated the results of their building archetype development process based on aggregated energy consumption (i.e. the total energy consumption of the building stock) or at the individual building level (Reinhart and Cerezo Davila 2016). On an aggregate level, archetypes predicted building stock annual energy consumption or energy use intensity with a relative error between 4% and 21% and a mean value of 11% error. The accuracy of the energy predictions for specific buildings decreased significantly, with errors for individual building energy prediction reaching as high as 99% for certain cases. Archetypes found in the literature predict the annual energy consumption of a building stock well, but break down at the building scale.

Some more recent studies have shown that having access to monitored data for a small subset of the building stock can significantly increase the accuracy of the building archetypes. Sokol et al. calibrated archetypes with annual and monthly energy consumption to improve the prediction accuracy of the stock model (Sokol et al. 2016). The most significant improvement was due to the annual calibration, which decreased the model error from 41% to 14%. Further calibrating on a monthly basis resulted in a relative error of 9% for the annual energy consumption prediction. The main benefit for the monthly calibration was much improved monthly energy prediction, which would be essential for studying seasonal effects such as the peak energy usage. For peak load prediction of a building stock, archetypes have been shown to reach errors up to 40% between the archetype result and actual peak load (Heiple and Sailor 2008).

### **2.2.3.4 Archetype summary**

Building archetypes have been successfully applied to predict the annual energy consumption of building stocks of varying scales, including the urban, regional and national levels. The number of archetypes used to characterize each stock depends greatly on the nature of the buildings and the selected geographic region. Increasing the number of segments and the categories per segment greatly increases the number of potential archetypes in an exponential manner. Archetypes perform well for annual energy consumption prediction, achieving relative errors on the order of 11% on average for the whole building stock.

The gaps in the literature regarding building archetypes can be summarized by the following points:

- Building archetypes are unable to accurately predict peak loads for building stocks (Heiple and Sailor 2008)
- Archetypes are limited by the information that is available for a building stock (Reinhart and Cerezo Davila 2016)
- Data sets for at least a subset of a building stock can improve archetype accuracy (Sokol et al. 2016)
- Archetypes are generally inaccurate for prediction of a single building's energy consumption within the building stock (Reinhart and Cerezo Davila 2016)

#### **2.2.4 Examples of bottom-up white-box stock models based on physics simulation**

A summary of recent bottom-up white-box building stock energy models based on physics simulation is provided in Table 2.4. The features and limitations of each model are presented. Each model falls in the Q4 quadrant, though some are technically hybrid models, such as the Scout model (Harris et al. 2016).

Q4 stock models tend to fall in one of two categories: 1) represent 100% of buildings in the studied building stock, or 2) represent a sample of buildings. The first portion are intended for urban applications and appear to be generally limited to a sample of 10,000 buildings or less and target both commercial and residential buildings, e.g. CityBES (Hong et al. 2016). The second portion can represent larger building stocks, such as National stocks, but use a very small portion of the stock as a sample, i.e. less than 1% of buildings, such as the TREES model (Taniguchi-Matsuoka et al. 2020).



Table 2.4 Summary of recent bottom-up white-box building stock energy models (Q4 models)

Model name	Stock	Stock size <sup>1</sup>	Building sample count (%)	Features	Limitations	Market	Ref.
UMI	Urban, user-defined	~30	30 (100%)	Flexibility, customized input of an urban building stock, multiple end-use applications. Static model.	The accuracy of the stock model is not clear, uses a “shoe-box” simplification to approximate building geometry and exposure.	Commercial, residential	(Reinhart et al. 2013)
CityBES	Urban, user-defined	10,000	10,000 (100%)	Flexibility, customized input of an urban building stock, multiple end-use applications. Static model.	Like UMI, the accuracy of CityBES is unclear. Size of stock is limited due to computational requirements.	Commercial, residential	(Hong et al. 2016)
Scout	National (USA)	123 million households	N/A	Impact evaluation of energy conservation measures on energy and CO <sub>2</sub> emissions. Dynamic model.	Scout appears to be closer in functionality to a top-down model (Q1), ideal for long-term prediction. Difficult to assess the accuracy of the model on smaller stock samples.	Commercial, residential	(Harris et al. 2016; Jared Langevin, Harris, and Reyna 2019)
ResStock	National (USA)	123 million households	350,000 (0.28%)	Visualization tools, national baseline. Energy sources and end-uses by state. Static model.	Certain house combinations result in significant discrepancies, overall error of the stock model generally within $\pm 20\%$ of validation data.	Residential	(Wilson et al. 2017)
AutoBEM	Urban, user-defined	130,000	130,000 (100%)	Flexibility, uses a number of imaging techniques to build 3D maps of urban settings. Static model.	The tool provided online is difficult to assess and the energy intensity for certain buildings appears quite elevated.	Commercial, residential	(New et al. 2018; ORNL 2021)
Synthetic building stock tool	National (Switzerland)	1.6 million households	10,000 (0.6%)	Auto generation of stock characteristic distributions. Static model.	Monthly energy consumption prevents determination of peak energy use. Limited to 10,000 buildings.	Residential	(Nägeli et al. 2018)
TREES	National (Japan)	53 million	16,000 (0.03%)	Detailed occupancy and house characteristics. Static model.	Very limited sample of buildings used to represent the stock, limited accuracy and simplistic thermal network representation.	Residential	(Taniguchi-Matsuoka et al. 2020)

<sup>1</sup> Stock size is based on the example use case provided by the authors of the tool.

The examples illustrate that building stock energy models that rely on physics simulations are limited either by scope or by sample size. Much like building archetypes, accuracy is a recurring issue with stock models, as the accuracy is either not reported by tool authors, the model has significant discrepancy with stock measured data, or the tool requires additional calibration.

### **2.2.5 Challenges in building stock energy modeling**

As with many fields, a number of unanswered questions remain in building stock energy modeling. Booth et al. (2012) identified five main limitations with respect to the development of BSEM:

1. Accuracy
2. Data collection
3. Computational time
4. Decision-making
5. Flexibility

These five categories are addressed in the following subsections.

#### **2.2.5.1 Accuracy**

Accuracy is one of the primary concerns of any energy modeler, and in the case of stock modeling can be of particular concern as errors can propagate from the building-level to the stock-level in unpredictable ways. The accuracy of building archetypes is discussed in Section 2.2.3.3, a subject addressed in detail by Reinhart and Davila (2016). Stock energy prediction using archetypes can be accurate with a reasonable margin while errors on the building-level can reach as high as 99%. Errors for peak load prediction in urban settings can reach 40% using archetypes (Heiple and Sailor 2008). For other physics simulation-based models, accuracy is often not addressed at all. Energy models for long term prediction have no data for validation for future years. Work related to stock model uncertainty appears to be ongoing in IEA Annex 70 and has yet to be addressed in the literature (J. Langevin et al. 2020).

#### **2.2.5.2 Data collection**

The data required for a building stock energy model varies according to the approach taken by the modeler. For statistics-based models, real energy consumption data is required to develop

regression models or other inverse modeling techniques. For engineering models the data collected is more often stock-related data, such as the synthetic stock data generated by Nägeli et al. (2018). The segmentation and characterisation processes used for building archetypes are another example of data collection (Sokol et al. 2016). The danger for physics simulation stock models is to represent buildings that do not exist in the stock by combining building parameters that are not consistent with the real stock. Access to relevant, real data remains one of the main limitations of stock models.

### **2.2.5.3 Computational time**

The study by Booth et al. (2012) was performed in 2012, and since then there have been significant advancements in computer components, computational resources and cloud computing. Nevertheless, several of the stock models identified in Table 2.4 mention computational resources as a limitation related to the sample size or total sample allowed by the stock tool, without mentioning the impact that a reduced sample size has on the overall model accuracy. Some national stock models sample as little as 0.03% of the stock (Taniguchi-Matsuoka et al. 2020).

### **2.2.5.4 Decision-making**

Building stock modeling tools are designed for a purpose, often to address one of the industry needs identified in Section 2.2.1. The output of a model is designed to serve that purpose but often the effect of the uncertainty inherent to the model is not stated or evaluated by the tool creators. The outcome is therefore uncertain, and a user of the tool has no method of evaluating the impact of the uncertainty on the decisions made using the tool. No studies addressing uncertainty propagation and the impact on decision-making using model outcomes have been performed in the literature (Booth, Choudhary, and Spiegelhalter 2012; J. Langevin et al. 2020).

### **2.2.5.5 Flexibility**

The flexibility of a building stock model pertains to the capability of the model to adapt to technological advancements over time. For example, if fossil fuels are eliminated from a particular market, or occupancy behaviour is drastically altered via aggressive time-of-use tariffs. Statistics-based models are often less flexible because they are based on historical data, i.e. based on the way

things were in the past. Flexibility is one of the advantages of engineering-based or white-box models, as simulation tools tend to be adaptable to new technologies.

## **2.3 Conclusion**

The literature review provides an overview of the challenges associated with residential building stock energy modeling (BSEM). Additional literature is presented in each of the articles of this thesis, where specific methods are evaluated in more detail. In summary, the use of a BSEM is integral to a variety of industry needs and yet there are no practical BSEMs available for use in Canada. The key issues facing researchers in BSEM development can be described as follows:

- 1) Data collection is an issue often associated with building archetype development, though the segmentation and characterisation processes are interesting techniques to describe the key building parameters affecting energy consumption for any building stock energy model.
- 2) Computational resources remain a factor for stock modeling, though the limitation of sample size and the impact that it has on overall accuracy of a stock model remains unexplored.
- 3) Overall accuracy of building stock energy models remains somewhat of a mystery, as the accuracy of many models is either not divulged or unknown to the tool authors, and stock models based on building archetypes tend to vary widely in accuracy depending on the approach used.
- 4) Flexibility is one of the key desired features of industry for BSEM, as the ability to evaluate technological changes and the impact on energy and peak electricity usage are integral to the evaluation of GHG emissions calculations and the impact that shifting fossil fuels to electricity has on the electric grid, among other factors. Commonly used techniques, such as building archetypes and black-box (statistical) methods, have difficulty assessing technological evolution.

## CHAPTER 3 OBJECTIVES AND THESIS ORGANISATION

The literature review identified a number of key limitations affecting the development of building stock energy models. This thesis aims to address some of those limitations and to describe the development of a new single-family dwelling building stock energy model for the province of Québec, Canada. The overall objectives and the structure of the thesis are presented in the following two sections.

### 3.1 Thesis objectives

The objective of this thesis is to address the *data sourcing*, *accuracy*, *computational resources*, and *flexibility* limitations inherent to the development of a building stock energy model and discussed in Section 2.2.5 of the thesis, while providing a new building stock energy modeling tool for single-family dwellings. More specifically, this research can be described in terms of three primary objectives:

- Objective 1: Develop a stochastic virtual smart meter data set for single-family dwellings, designed for developing predictive models that can characterize building stock parameters in the absence of real data
- Objective 2: Evaluate linear discriminant analysis as a supervised machine learning classification algorithm for the prediction of building parameters from smart meter data
- Objective 3: Develop a new bottom-up white-box building stock energy model for single-family dwellings

Each of the objectives above addresses one or more of four typical limitations of building stock energy modeling identified by Booth et al. (2012). *Decision-making* is not addressed in this study as it requires a completed building stock model to evaluate uncertainty propagation in stock modeling and other related aspects.

### 3.2 Main structure of the thesis

The research related to this thesis resulted in five publications: two conference papers with international scientific committees and three peer-reviewed journal articles. The first conference paper was presented at the eSim 2018 conference and discussed the overall methodology for stochastically generating virtual smart meter data, i.e. electricity consumption generated using building energy simulations, for future use in classification studies (Neale, Kummert, and Bernier 2018). The second conference paper included a preliminary evaluation of linear discriminant analysis as a technique for predicting building parameters from metered data, which was published in the international conference BSim 2019 in Rome, Italy (Neale, Kummert, and Bernier 2019). The second conference paper was among six finalists for the best student paper of the conference.

Journal article 1 (J1), presented in **Chapter 4**, builds upon and completes the methodology from the first conference paper. Published in the *Journal of Building Performance Simulation*, article J1 titled “Development of a stochastic virtual smart meter data set for a residential building stock-methodology and sample data” presents a methodology and sample virtual smart meter data set ideal for evaluation of classification methods using electricity smart meter data (Neale, Kummert, and Bernier 2020a). Very few smart meter data sets with known building parameters exist, and none are designed specifically for classification studies. In addition, article J1 illustrates that modeling a large stock of 200,000 buildings is feasible. The data set can be used to train predictive models that can read electricity smart meter data and establish property distributions for building stocks.

Journal article 2 (J2) uses the virtual smart meter data set developed in J1 to evaluate the potential of predictive models developed using linear discriminant analysis. The article J2, presented in **Chapter 5** and titled “Discriminant analysis classification of residential electricity smart meter data”, illustrates the effectiveness of LDA as a prediction tool for building parameters and guides others on the number of features and size of the data set appropriate for future classification studies. The classification accuracy is determined for a variety of cases and model development time is discussed. Article J2 was published in the journal *Energy and Buildings* in December 2021.

Journal article 3 (J3) presents a complete building stock energy model for a provincial single-family dwelling building stock. Presented in **Chapter 6** and titled “Development of a new bottom-up white-box building stock energy model for single-family dwellings”, article J3 describes a new building stock energy model for homes in the province of Québec, Canada. The methodology used to produce the virtual smart meter data is extended to model the entire single-family dwelling building stock, including natural gas, heating oil and wood-based energy systems. The precision of the developed model is evaluated by comparing it to provincial energy data for 30 different energy consumption totals. The impact of the modeled sample size compared to the overall stock size is discussed. Two case study scenarios are compared to the status quo building stock to evaluate the impact on provincial greenhouse gas emissions, energy consumption and peak electricity use. Article J3 was submitted to the *Journal of Building Performance Simulation* in November 2021.

Finally, **Chapter 7** presents an overall discussion of the results of the thesis, followed by some conclusions and recommendations for future work.

## **CHAPTER 4      ARTICLE 1: DEVELOPMENT OF A STOCHASTIC VIRTUAL SMART METER DATA SET FOR A RESIDENTIAL BUILDING STOCK - METHODOLOGY AND SAMPLE DATA**

Neale, Adam, Michaël Kummert, and Michel Bernier. 2020. *Journal of Building Performance Simulation* 13 (5): 583–605.

### **4.1 Abstract**

Existing electricity smart meter data sets lack sufficient details on building parameters to evaluate the impact that home characteristics can have on electricity consumption. An extensive, open-source virtual smart meter (VSM) data set with corresponding building characteristics is provided. The methodology used to develop the VSM data is presented in detail. The data set consists of a variety of homes representative of a subset of the Canadian single-family home building stock. The building characteristics cover a wide range of values that are based on probability distributions developed using a segmentation and characterization process. The resulting framework and VSM data set can be used by researchers to develop classification models, verify load disaggregation algorithms, and for a variety of other purposes.

### **4.2 Introduction**

Estimating the energy use of multiple buildings, such as at the city- or regional-scale, is commonly referred to as building stock energy modeling (BSEM) or urban building energy modeling (UBEM). BSEM can be accomplished with a number of well-documented techniques (Swan and Ugursal 2009). One such technique makes use of building archetypes to simplify the modeling approach through a process of segmentation and characterisation of the building stock characteristics (Reinhart and Cerezo Davila 2016).

Building archetypes have been developed for numerous cities and countries across the globe. One of the key issues facing archetype development is the lack of reliable and accurate information on the buildings (Booth, Choudhary, and Spiegelhalter 2012). The accuracy of the building stock model depends on the quantity and quality of available data. In addition, data sources vary widely



at municipal, regional and federal levels, which makes it difficult to apply a single methodology across all building stocks.

This work is part of a broader study to provide a new methodology for using electricity smart meter data as a data source for building stock characterisation and segmentation. A key component of this research consists in the development of classification models using machine learning on smart meter data sets. With a sufficiently large smart meter data set with known building characteristics, classification models could be developed and used to determine the characteristics from smart meter data for a variety of regions. However, this approach depends on two factors: 1) sufficient smart meter data availability, and 2) building characteristics that are associated with each set of smart meter data. To discuss the possibility of classification modeling, smart meter data availability must first be addressed.

#### **4.2.1 Smart meter data**

As of 2017, over 770 million electricity smart meters have been installed globally (IEA 2019). This amount has been steadily increasing in recent years, in particular due to significant interest in the Asia-Pacific region. There are 79 million meters in the United States of America alone (IEA 2019). The global market in terms of installed units increased by over 12% from 2016 to 2017 as countries seek to convert their building stock to new metering technologies (IEA 2017).

In the province of Québec, Canada, there are over 3.7 million electricity smart meters (Hydro-Québec 2016) currently installed, primarily for the 3.6 million household residential market (2016 data, Natural Resources Canada 2019a). These meters collect data at 15-minute intervals and are mostly used for billing purposes (Hydro-Québec 2012).

Smart meters are very prevalent and could present an interesting opportunity to extract information about residential energy consumption. While numerous applications are possible for analysing smart meter data, the authors focus primarily on developing classification models of residential smart meter data based on known building characteristics.

### 4.2.2 Supervised machine learning classification of smart meter data

While the purpose of this paper is not to discuss classification model development, a brief mention is made here for context. Supervised machine learning classification can be performed on electricity consumption data to estimate building parameters, so long as the classification algorithm has sufficient data to train the model. Unsupervised machine learning is typically unsuited for this process, as there is limited information on a building that can be extracted from electricity consumption data without training the model.

Some classification studies have been performed on real smart meter data (see e.g. Beckel et al. 2014; Carroll et al. 2018; Neale et al. 2019). To classify smart meter data requires:

- Electricity consumption at a given sampling rate (classification predictor);
- Known building parameters (classification response).
- An appropriate classification algorithm (e.g. discriminant analysis).

Since the objective of the authors is to use classification on smart meter data, it is important to discuss available smart meter data sets.

### 4.2.3 Electricity smart meter data available for classification studies

There are very few open-source residential electricity meter data sets with known building characteristics. Those that do exist have very little information on the buildings. Utilities are often reluctant to share the meter data due to privacy reasons. Of the electricity data sets that were found, the works were divided into two common themes: 1) smart meter data for occupant behaviour analysis, and 2) high-resolution submetered electricity for load disaggregation studies. Data sets were researched based on their location, number of homes in the sample, sampling frequency, trial duration and relevant building characteristics, which are presented in Table 4.1.

Table 4.1 Summary of open-source residential smart meter data sets with relevant building information

Data set	Loc.	# homes	Sampling period	Trial duration	Building information	Ref.
<b>Smart meter data sets</b>						
CER electricity customer behaviour trial	Ireland	4232	30 minutes	1.5 years	Occupant social data, appliance use, some building geometry information.	CER (2012)
PNNL GridWise Demonstration Project	USA	112	15 minutes	~1 year	Occupant surveys	Hammerstrom, Ambrosio et al. (2007), Hammerstrom, Brous et al. (2007)
<b>Load disaggregation data sets</b>						
UMass Smart* Home Dataset (2017 release)	USA	7	1 minute	Varies (1-2 years)	Weather data, very detailed submetered electricity consumption.	Barker et al. (2012)
UMass Smart* Microgrid Dataset	USA	443	1 minute	24 hours	Electricity consumption only.	Barker et al. (2012)
REFIT smart home dataset	UK	20	Mixed	2 years	Building occupant survey data, complete building description, high-resolution appliance electricity use.	Murray et al (2017).
Almanac of Minutely Power dataset 2	Canada	1	1 minute	2+ years	Building geometry	Makonin et al. (2016)
Dutch Residential Energy Dataset	The Netherlands	1	1 second	6 months	Building geometry, occupancy, appliances, indoor temperature	(Nambi et al. 2015)
ECO dataset	Switzerland	6	1 second	8 months	Occupancy data	Kleiminger et al. (2015)
Carleton high-resolution electricity data set – Study 1	Canada	12	1 minute	~14 months	Type, vintage, building surface area, number of occupants	Saldanha & Beausoleil-Morrison (2012)
Carleton high-resolution electricity data set – Study 2	Canada	23 <sup>1</sup>	1 minute	~1 year	Type, vintage, building surface area, number of occupants	Johnson & Beausoleil-Morrison (2017)

<sup>1</sup> Includes the 12 homes from the study by Saldanha & Beausoleil-Morrison (2012).

The smart meter data sets are characterized by a relatively high number of homes and a monitoring period exceeding 1 year. They focused on occupancy behaviour and contained little or no details on the characteristics of the buildings. The Irish Commission for Energy Regulation (CER) electricity customer behaviour trial (CER 2012) is the best example of open-source smart meter data with over 4000 single-family homes. Relevant building characteristics include the type of building (detached, semi-detached), building floor area and number of occupants. The CER data

set does not include any heating or cooling electricity use as these Irish homes tend to have no cooling system and have non-electric heating (Beckel et al. 2014).

Load disaggregation data sets are included in this review since they typically contain more detailed information on the building's characteristics and the high-resolution electricity consumption data can be aggregated. Due to the sheer quantity of data and measurement points, load disaggregation studies typically limit the scope of the monitoring to a smaller number of homes and/or shorter monitoring period. There is therefore little diversity in house types for these types of studies. The UMass Smart Microgrid Dataset has the highest sample of buildings with 443 homes, however it only contains the electricity consumption data and no building information (Barker et al. 2012).

The available data sets are therefore inadequate for developing classification models based on the electricity consumption at sampling rates representative of electricity smart meter data. There is either too limited information on the building to classify data, too few homes to represent the diversity of the building stock, or too little data to represent a full year of electricity consumption.

### **4.3 Objectives**

The principal objective of the authors is to estimate building characteristics from anonymous electricity smart meter data, an approach previously described in Neale et al. (2018). While this would normally be accomplished by training classification models using a smart meter data set with known building parameters, no such data sets are extensive enough to perform that task at this stage. The main objective therefore is to develop a virtual smart meter (VSM) data set with known building characteristics using batched building energy simulations. The VSM data set will be used to develop a clear link between building characteristics and electricity consumption at actual smart meter data resolution.

The single-family home market is targeted as it is a significant portion of the residential building stock in Canada, representing 16.5% of secondary energy use (2017 data, NRCan 2020). Expansion to other markets is possible in future work. Given the need to train classification models for many different building characteristics, the building characteristics must be generated in a way to be as

representative as possible of the building stock, to minimize combinations of parameters that are not likely to appear in the chosen market.

The developed framework must therefore have the capability to:

- Generate single-family homes using building energy simulation;
- Use building characteristics and conditions typical of the chosen market using probability distributions, wherever possible;
- Determine the electricity consumption of the home, effectively producing a “virtual smart meter data profile”;
- Link the building characteristics to the virtual smart meter data profile for each generated building;
- Generate a large quantity of homes (e.g. 100,000+ homes) to best cover the range of possible building parameters for classification model development, i.e. via batch simulation.

## 4.4 Methodology

The methodology used to develop the framework and produce the VSM data set for an example building stock is presented in a way that can facilitate applying it to other building stocks. A sample set of VSM data with the corresponding building characteristics is provided as supplementary material with the paper. This open-source data set can be used to test classification algorithms and study the impact of various building parameters and occupancy profiles on electricity use at typical smart meter sampling periods.

The framework concept is divided into two main components illustrated in Figure 4.1: 1) *Generator*, which is the component that produces the VSM data profiles, and 2) *Classifier*, which is the classification model module. While the *Generator* component is the focus of this paper, it is relevant to present the *Classifier* module in part to justify certain choices made in the development of the former.

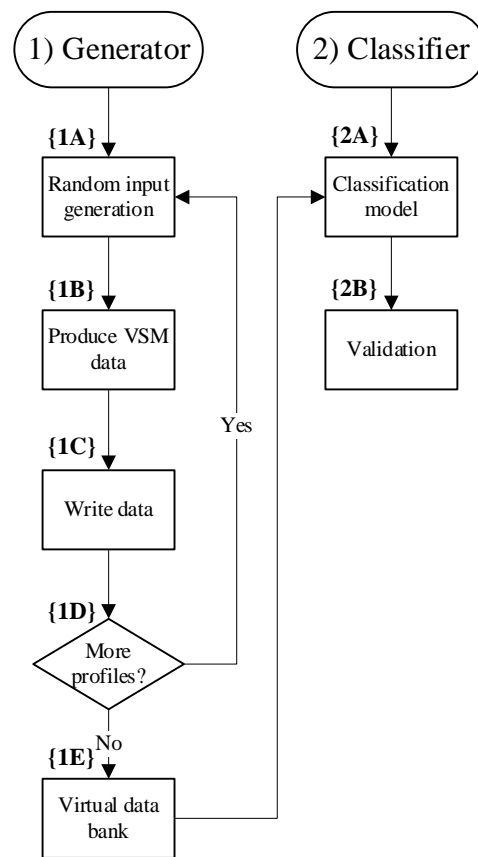


Figure 4.1 Virtual smart meter data generation (Generator) and classification (Classifier) processes

The *Generator* component consists of five parts, described as {1A} to {1E} in Figure 4.1. First, the building parameters are generated randomly {1A} to ensure a unique building is generated upon each iteration of the model. This is accomplished by generating building characteristics according to predetermined probability distributions. A building simulation is then performed {1B} and the data are output in a specific format corresponding to typical smart meter data {1C}. The module then determines whether additional profiles are required {1D} and the process would continue. If the data generation is complete, the VSM profiles are compiled into a single data bank for future use {1E}.

The *Classifier* component then reads the data bank and develops a classification model based on the smart meter data and known building parameters {2A}. The result would then be validated using real smart meter data for the targeted building stock, i.e. the building characteristics for the

real smart meter data would be compared with predicted values from the classification process {2B}.

## **4.5 Virtual smart meter data generation**

Virtual smart meter (VSM) data is defined here as electricity consumption data generated using a physics-based building energy model at 15-minute intervals, though other sampling rates are possible. The sampling rate selected for the virtual data should match the one used by the local electricity distributor.

VSM data is intended to replicate the electricity consumption for single-family homes of various compositions. While the methodology behind the model can be applied to any region, the characteristics selected are representative of homes found in the province of Québec, Canada.

### **4.5.1 Development of a virtual smart meter framework**

As illustrated in Figure 4.1, the purpose of the developed framework is to stochastically generate a set of building parameters and then execute a building energy simulation using those parameters. The framework produces a VSM data profile that consists of electricity consumption data (in kWh) at 15-minute intervals, which is 35 040 data per home for a year. Each data profile is paired with the building parameters that are used to generate it. This process is then repeated the desired number of times to create a large data set for classification. As an example, the authors typically generate 200,000 homes with corresponding inputs and electricity consumption, which is well over 7 billion data.

The single-family home building stock for the province of Québec, Canada, is characterized by a variety of detached, semi-detached and row houses that typically have a basement. The homes are usually either one- or two-storeys and have a variety of thermal envelope performance levels and occupancy characteristics. Housing density (and therefore housing types) varies across the province, as does the climate. The framework therefore must include the following features:

- Generating physical characteristics, such as size, shape, number of floors, for a home in the province of Québec;

- Using a weather data file that represents the climate for each region of the province of Québec;
- Occupying the virtual home with realistic occupants and internal loads and simulating their demands;
- Producing the annual electricity use for the heating, cooling, lighting, appliances and domestic hot water loads of the home;
- Repeating the process a large number of times with statistically representative inputs each time.

In order to generate a set of housing electricity consumption that is realistic, a framework had to be developed that could generate parameters that are representative for the chosen building stock and that correctly impact the house's energy use.

#### **4.5.2 VSM framework details**

The proposed framework consists of two main components: 1) a manager, which generates the building characteristics and starts a building energy simulation automatically, and 2) the building energy model, which generates the VSM data set given the set of selected parameters. The manager is an essential part of the process of automating the generation of building parameters and batch building simulations required to produce a significant set of VSM data. The logic behind the manager-building model interaction is illustrated in Figure 4.2.



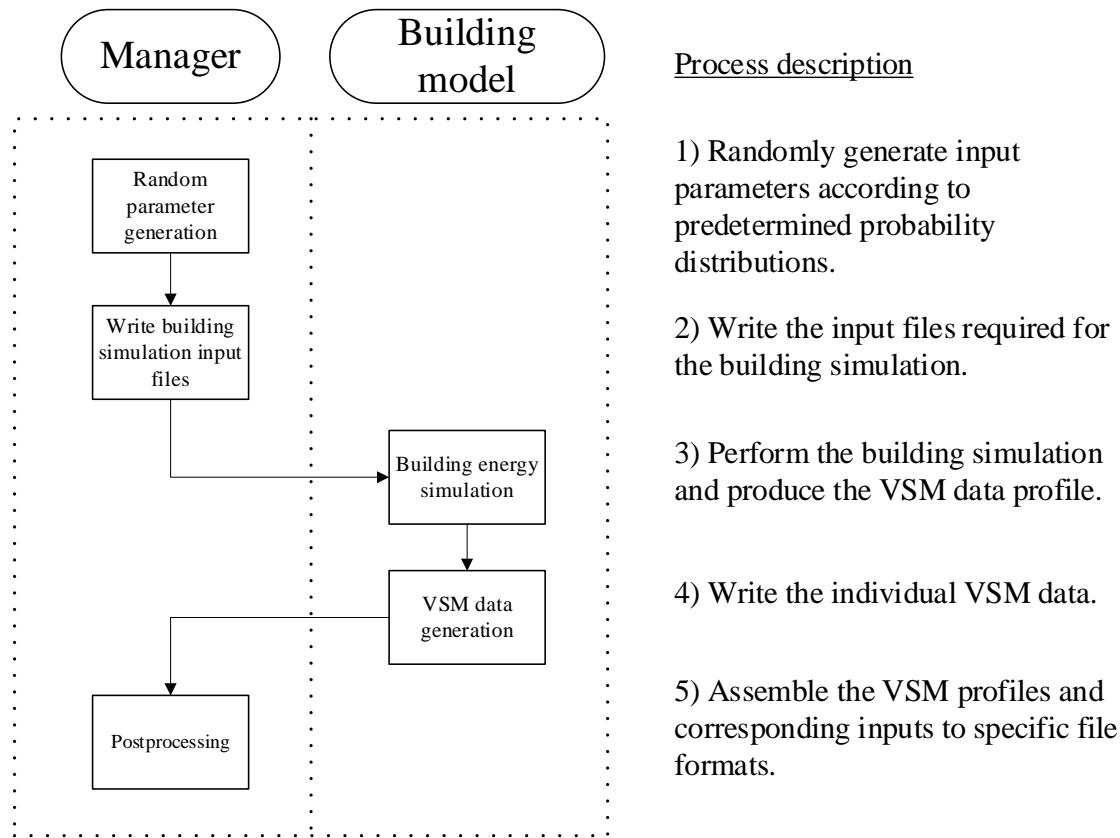


Figure 4.2 Proposed framework including a manager and building energy model

The manager is implemented in the Matlab environment (Mathworks Inc. 2018a). The building model used is the Type 56 implemented within the TRNSYS environment (Klein et al. 2017). This model takes into account thermal mass to perform a transient simulation. Other dynamic simulation environments and building simulation programs could work equally well for the task. The building modeling approach for determining the building geometry, heating and cooling systems, lighting, appliance, domestic hot water and air infiltration loads are discussed in the following sections of the paper.

### 4.5.3 Building geometry

The building is represented as a single-zone rectangular prism. By default, the building is oriented with the south-facing wall as the “front” of the home, i.e. the façade that is facing the street. If the house type is semi-detached or a row house, one or both of the east- and west-facing walls are

considered adjacent to another dwelling and not exposed to outdoor conditions. Houses can be one- or two-storeys. The heated surface area of a home is considered the sum of the floor areas including a basement. The building rotation is specified in order to determine the solar gains for each surface. The aspect ratio of the building's depth to width determines the area of each external surface. The height of each storey is considered constant (2.4 m).

External walls and roof surfaces are insulated with specified thermal resistance values. Windows are evenly distributed on all external envelope surfaces based on a window-to-wall ratio. Walls that are shared between more than one dwelling do not contain any windows, such as in semi-detached or row houses.

The building model includes a basement that is primarily below ground level, where the foundation and slab are insulated with a specified thermal resistance value. The ground temperature is modelled with a sinusoidal external boundary condition described in Equation (4.1).

$$T_{ground} = C_0 + C_1 \cos(C_2 t + C_3) \quad (4.1)$$

where  $T_{ground}$  is the ground temperature on the outer surface of the building envelope ( $^{\circ}\text{C}$ ),  $C_i$  are constants, and  $t$  is time (h). Constants are specific to the region where the house is located and whether the external surface is a foundation wall or beneath a slab. Coefficients in equation 1 were pre-calculated for each region and determined based on a 3-D finite-difference type simulation to predict heat transfer in basement walls and slabs.

#### 4.5.4 Heating systems

The heating system for a home can consist of one of three possible system configurations (note: symbols for equations appear at the end of this section):

- (1) Electric element heating, i.e. electric baseboards or an electric furnace. The heating load for the home is calculated and a constant coefficient of performance (COP) is applied to calculate the electric heating requirement.  $COP_{electric} = 1.0$ ,  $PF = 1.0$ .

$$E_{Heat,elec} = \frac{PF \times Q_{Heat,load} \times \Delta t}{COP_{electric}} \quad (4.2)$$

- (2) Non-electric heating, i.e. a natural gas or heating oil furnace. The heating load for the home is calculated and a parasitic load is applied for electric subsystems and applied in the form of a power fraction (PF).  $COP_{electric} = 1.0$ ,  $PF = 0.05$ .

$$E_{Heat,elec} = \frac{PF \times Q_{Heat,load} \times 0.25}{COP_{electric}} \quad (4.3)$$

- (3) Air-source heat pump (ASHP) with an auxiliary (AUX) heating system. In this case, the heating system transitions between the ASHP and AUX systems based on the outdoor temperature. The usage fraction for the ASHP ( $F_{ASHP}$ ) and AUX ( $F_{AUX}$ ) are expressed using the following heuristic relationships:

For  $T_{outdoor} > -5 \text{ }^{\circ}\text{C}$

$$F_{AUX} = 0$$

$$F_{ASHP} = 1.0$$

For  $-12 \text{ }^{\circ}\text{C} \leq T_{outdoor} \leq -5 \text{ }^{\circ}\text{C}$

$$F_{AUX} = 0.1429 \times T_{outdoor} + 1.7143$$

$$F_{ASHP} = 1 - F_{AUX}$$

For  $T_{outdoor} < -12 \text{ }^{\circ}\text{C}$

$$F_{AUX} = 1.0$$

$$F_{ASHP} = 0$$

The COP for the ASHP is determined as a function of outdoor air-temperature, as depicted in Equation (4.4).

$$COP_{ASHP} = 0.0585 \times T_{outdoor} + 3.115 \quad (4.4)$$

The overall heating system electricity usage for a heat pump system with auxiliary heating for a given time step is expressed using Equation (4.5).

$$E_{Heat,elec} = \frac{PF \times F_{AUX} \times Q_{Heat,load} \times \Delta t}{COP_{AUX}} + \frac{F_{ASHP} \times Q_{Heat,load} \times \Delta t}{COP_{ASHP}} \quad (4.5)$$

Description for symbols used in Equations (4.2) through (4.5):

- $\Delta t$  is the time step in the simulation (0.25 h).
- $COP_{AUX}$  is the COP of the auxiliary heating system.
- $COP_{ASHP}$  is the COP of the air-source heat pump for the given time step.
- $E_{Heat,elec}$  is the amount of electricity required to heat the home for the given time step (kWh).
- $F_{AUX}$  is the usage fraction for an auxiliary (AUX) heating system.
- $F_{ASHP}$  is the usage fraction of an air-source heat pump (ASHP).
- $PF$  is the power fraction for non-electric equipment.
- $Q_{Heat,load}$  is the heating load calculated by the building model (kW).
- $T_{outdoor}$  is the outdoor dry bulb temperature ( $^{\circ}\text{C}$ ).

#### 4.5.5 Cooling systems

The electricity use for cooling a home is determined based on the presence of a cooling system. If no cooling system is present, the cooling electricity is zero. If an air-conditioner or a reversible heat pump exists, the cooling electricity use is modelled using Equation (4.6).

$$E_{Cool,elec} = \frac{Q_{Cool,load} \times \Delta t}{COP_{ASHP}} \quad (4.6)$$

where  $E_{Cool,elec}$  is the amount of electricity required for cooling for the given time step (kWh),  $\Delta t$  is the duration of a time step (0.25 h), and  $COP_{ASHP}$  is the coefficient of performance calculated as a function of outdoor temperature, as expressed in Equation (4.4).

#### **4.5.6 Lighting internal gains**

Lighting is modelled as an internal heat gain corresponding to a heat source expressed in watts (W). Lighting heat gains to the surrounding environment are considered 57% radiative and 43% convective (ASHRAE 2013).

#### **4.5.7 Equipment internal gains**

Equipment internal gains are applied as an internal heat source expressed in watts (W). Heat gains to the surrounding environment are considered 30% radiative and 70% convective (ASHRAE 2013). Note that some equipment, such as swimming pool and spa pumps and heaters, are not typically installed within the home. While these devices are included in the total electricity consumption, they are not included as internal gains within the house.

#### **4.5.8 Domestic hot water**

Domestic hot water electricity use is determined using a hot water tank simulation using the TRNSYS software (Klein et al. 2017). The water heater model consists of a vertical cylindrical insulated tank that is equipped with master-slave heating elements, controlled with aquastats and with a volume equal to 266 L. Standby losses are determined by assuming a constant ambient air temperature and constant thermal resistance of the tank. Electricity use is calculated based on the activation of the heating elements due to temperature changes in the hot water tank subsequent to hot water draws.

#### **4.5.9 Air infiltration**

The Sherman-Grimsrud infiltration model is used to represent air infiltration, which determines the air changes per hour (ACH) as a function of the indoor and outdoor temperatures, the leakage area of the building envelope, the wind speed given the height of the building, and the pressure differences due to stack effect (Sherman and Grimsrud 1980).

## 4.6 Building parameters

A residential single-family building can be described using a variety of deterministic and probabilistic parameters. The number of virtual buildings required in the data set depends on the parameters being varied and the discretization of each for classification purposes. In other words, each characteristic is divided into discrete “bins” that are subsequently used for classification model development. Parameters are generally divided into 2 to 5 bins to ease the classification model development (Neale et al. 2019). In brief, if the goal is to correctly predict the category for a given building parameter, the likelihood of a correct prediction increases with fewer categories. The exactitude of the prediction is up to the classification modeler and the desired accuracy of the developed model.

A general description of the building parameters is provided in Table 4.2. The impact of the parameter on the building energy model is presented. The information available for each building stock will vary, though some examples of potential data sources are provided for each building parameter. Examples of categories for each characteristic are also provided in Table 4.2.

Table 4.2 Model inputs and potential data sources

Property	Impact on model	Potential data sources	Categories/bins <sup>1</sup>
<b>Location</b>	Distribution of building types, climate determines heating/cooling degree days	National census data, national energy use databases	Number of locations depends on region studied. User choice. (7)*
<b>Building type</b>	Determines building geometry	National census data, building energy surveys, municipal tax evaluation data	Single-detached, semi-detached, row are the usual categories. Semi-detached and row can be combined. (2-3)
<b>Shape</b>	Determines building geometry	Building surveys, national studies, engineering knowledge of construction practices	Aspect ratios, e.g. 0.8, 1.0, 1.3. User choice. (3-5)
<b>Rotation</b>	Determines building geometry, solar gains	Engineering knowledge, map data	90° rotation increments. (4)
<b># Floors</b>	Determines building geometry	Building surveys, national studies, engineering knowledge of construction practices	One or two floors. (2)

Property	Impact on model	Potential data sources	Categories/bins <sup>1</sup>
<b>Wall construction</b>	Determines building envelope thermal performance	Building surveys, national studies, engineering knowledge of construction practices	Wall thermal resistance levels. (4-5)
<b>Roof construction</b>	Determines building envelope thermal performance	Building surveys, national studies, engineering knowledge of construction practices	Roof thermal resistance levels. (4-6)
<b>Foundation construction</b>	Determines building envelope thermal performance	Building surveys, national studies, engineering knowledge of construction practices	Foundation thermal resistance levels. (4)
<b>Infiltration</b>	Impacts heating and cooling demand	Building surveys, national energy code levels, measurement campaigns	Rates of infiltration (3-5)
<b>Window type</b>	Determines building envelope thermal performance	Building surveys, national studies, engineering knowledge of construction practices	Multiple variations of each type are possible, but are categorized by number of glazings (Single, double, triple.). (3)
<b>Window-to-wall ratio</b>	Determines building envelope thermal performance and building geometry	Building surveys, national studies, engineering knowledge of construction practices	Surface area ratio window:wall, e.g. 0.1, 0.2, etc. (3)
<b>Basement</b>	Determines building geometry	Building surveys, national studies, engineering knowledge of construction practices	Yes/no/crawl space. (1-3)
<b>Heating, ventilation and air conditioning</b>	Determines heating, cooling and ventilation electricity use based on energy demand and fresh air needs	Building surveys, national studies, national energy use databases, equipment distributors.	Electric baseboards, central air, heat recovery ventilator, air conditioning. (2-3) each for cooling, heating, ventilation.
<b>Setpoints</b>	Determines heating and cooling demand	National studies, national energy use databases	Thermostat setpoints for heating and cooling, e.g. heating: 21 °C, cooling: 25 °C. (1-3)
<b>Appliances</b>	Determines appliance electricity use	National studies, national energy use databases	Number and type of appliances (N/A) <sup>2</sup>
<b>Lighting</b>	Determines lighting electricity use	National studies, national energy use databases	Type of lighting, density. (N/A) <sup>2</sup>
<b>Occupants</b>	Impacts internal loads of the home	National census data, national studies	Number of occupants (5+)
<b>Occupancy</b>	Impacts internal loads of the home	Occupancy studies	Occupant schedule and activity schedule. (N/A) <sup>2</sup>

Property	Impact on model	Potential data sources	Categories/bins <sup>1</sup>
<b>DHW consumption</b>	Determines domestic hot water electricity use	National studies, research	Volume of hot water (N/A) <sup>2</sup>
<b>Pool/spa installations</b>	Determines electricity use due to a pool and/or spa installation	Equipment distributors, national studies, building surveys, building permits	Yes/no (2) each for pool and spa

<sup>1</sup> Values in parentheses represent the authors' recommended number of categories. Values with ( )\* depend on the chosen building stock.

<sup>2</sup> As described in the present paper, internal loads are calculated stochastically using an independent tool and provided as an input to the building energy model.

A segmentation and characterization process of the Québec building stock was performed in order to fill out the categories described in Table 4.2. Data sources included Canadian census data (Statistics Canada 2011; Statistics Canada 2016), the Canadian Survey of Household Energy Use (NRCan 2011), and the Energuide Housing Database (NRCan 2018). Similar information may not be available for all regions, in which case engineering knowledge can be sufficient to define an appropriate distribution for each parameter. Based on the available information, parameters were divided into four categories:

- Occupancy-driven internal loads;
- Uniform probability distributions, which describe a set of parameter categories with equal probability;
- Probability mass functions, which describe a set of unequal probabilities for a number of different categories for a given parameter.
- Fixed parameters, which were input to the model as constant values for all building simulations.

#### 4.6.1 Occupancy-driven internal loads

The number of occupants in a home can be established using a variety of data sources, which are discussed in subsequent sections of the paper. Of the building parameters described in Table 4.2 there are a number that are dependent on the occupancy of a home, i.e. the number of occupants that are at home and active in the house at any given time. Lighting, appliances and domestic hot water are primary examples of these loads.



To ensure a variety of stochastic occupancy behaviour that is directly tied to the internal loads of a home, the CREST thermal model (McKenna and Thomson 2016) is used to produce distinct internal load profiles for use in the *Generator* module. The CREST model is intended to produce daily profiles at 1-minute intervals, but was adapted to produce annual occupancy activity schedules with corresponding lighting, appliance and domestic hot water usage data at 15-minute intervals. The CREST model distinguishes between weekend and weekday behaviour but is limited to a maximum of 5 occupants. The presence of appliances (i.e. number of televisions and electronic devices), their usage (amount of time the device is operated based on occupancy), the number of occupants present at home, and their activity level are generated stochastically by the model based on real time-of-use probability tables. This ensures a wide variety of stochastic occupancy behaviour.

An example of a typical occupancy schedule with the corresponding domestic hot water usage is provided in Figure 4.3 for two full days. Occupancy (number of people in the house) and activity (number of people active in the house) are depicted for a 3-person household. Occupants are typically present during the night but not active, and absent during the day but active, though this varies from day to day. It should be noted that only occupants present in the house can influence the internal loads, and the activity level for an occupant not present is for descriptive purposes only. Domestic hot water draws occur when occupants are both present and active. Similar trends occur for lighting and appliance loads, though they are not depicted in Figure 4.3. Many appliances have constant or periodic electricity draws that are independent of occupancy, such as refrigerators. These are also represented in the CREST tool.

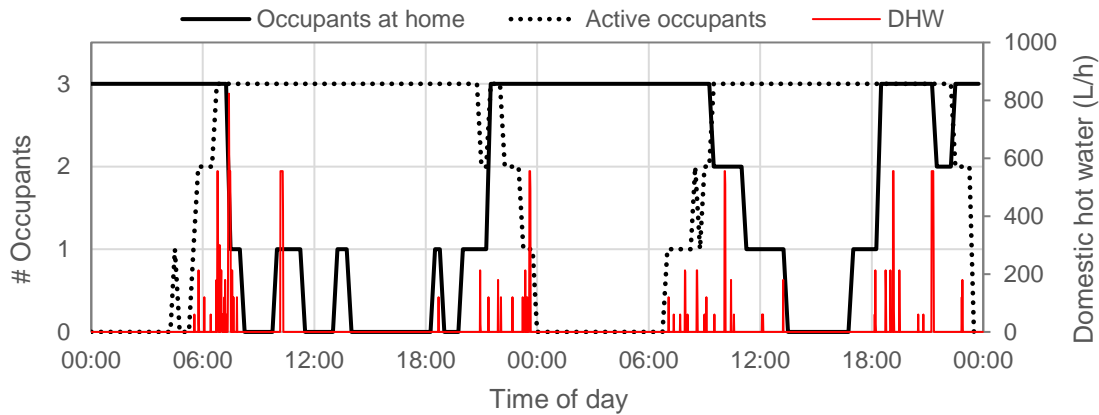


Figure 4.3 Example of an occupant activity schedule over a two-day period for a 3-occupant home

To generate a variety of occupant behaviours, 15 stochastic occupancy schedules for 1 to 5 occupants are generated, for a total of 75 occupancy profiles. The stochastic internal load profiles are provided with the VSM data set for context. If time-of-use data is available for the studied building stock, then more specific internal load profiles could be generated by updating the CREST probability tables or by implementing the internal loads in a different manner. Since the schedule and number of occupants are treated as separate inputs, it is possible to generate houses with similar physical characteristics but different occupancy.

In addition to the loads above, swimming pools and portable electric spas represent a non-negligible electric energy usage in the province of Québec. Swimming pools utilize between 4300 kWh and 7500 kWh annually depending on whether they are above- or below-ground, and whether they are heated (Hydro-Québec 2019a). The fraction of houses with swimming pools varies depending on the type of dwelling. Pools are considered operational from June 1<sup>st</sup> to October 31<sup>st</sup>. Portable electric spas were found to use between 4500 kWh and 5500 kWh annually, depending on the frequency of use (Hydro-Québec 2019b). Spas are considered operational all year long.

#### 4.6.2 Uniform probability distributions

A parameter described by a uniform probability distribution (UPD) has equal probability for all outcomes, as described in Equation (4.7).

$$P_{UPD}(A) = \frac{1}{k} \quad (4.7)$$

where  $P_{UPD}(A)$  is the probability for a building parameter  $A$  with a uniform probability distribution and  $k$  is the number of categories for that parameter.

UPD are typically applied in cases where no prior probability data could be found for the studied building stock. For example, data was not available for the Québec residential building stock to characterize the frequency at which each rotation value occurs, therefore equal probability was assigned to rotation values equal to  $0^\circ$ ,  $90^\circ$ ,  $180^\circ$  and  $270^\circ$ . While it is preferable to apply a probability mass function where data allows, a UPD will nevertheless cover the range of possible values for a parameter allowing for correct classification. The disadvantage of using uniform probability distributions without correlation to other parameters is that occasionally there may be combinations that do not exist in the building stock or that are overrepresented, using computational resources for the classification process for no real benefit. A list of UPD parameters and their corresponding number of categories and probabilities are presented in Table 4.3.

Table 4.3 Building parameters with uniform probability distributions

Parameter	Cate- gories	Category description	Uniform probability $P_{UPD}$	Notes
Rotation	4	Angle of rotation of the building with respect to true north. $0^\circ$ , $90^\circ$ , $180^\circ$ , $270^\circ$	0.250	GIS <sup>1</sup> data could help characterize this parameter.
Shape	5	Ratio of width to length for a house. 0.8, 0.9, 1.0, 1.1, 1.2	0.200	Semi-detached houses can have asymmetrical configurations and it is therefore important to represent both aspect ratio and building rotation. GIS <sup>1</sup> data could help characterize this parameter.

Parameter	Cate- gories	Category description	Uniform probability $P_{UPD}$	Notes
Window-to-wall ratio	3	Ratio of window surface area to aboveground vertical building envelope area. 0.1, 0.15, 0.2	0.333	Applied to all vertical building envelope surfaces that are above ground level that are not directly adjacent to another home, i.e. shared surfaces for semi-detached or row houses.
Occupancy profile	15	Stochastically generated occupancy profiles. 1 to 15.	0.067	Individually generated occupancy profiles. 15 profiles were generated for each occupant category, i.e. 15 profiles for 1 occupant, 15 profiles for 2 occupants, etc.
Adjacent building surfaces for semi-detached homes	4	Determines which external building surface is adjacent to a neighbouring home. “None”, “Both”, “East” or “West”.	0.250	Assuming that the front entrance of the building is facing south by default, the “east” and “west” terminology is adapted to describe which surface borders with a neighbour and therefore is not exposed to outdoor conditions.

<sup>1</sup> Geographic information systems

### 4.6.3 Probability mass functions

Probability mass functions (PMF) were established for a number of building parameters where statistical data was obtained. In many cases, sufficient data was available to establish dependence between one or more parameters, which is illustrated in Figure 4.4. The building location was selected first and subsequent parameters were chosen as a function of the location.

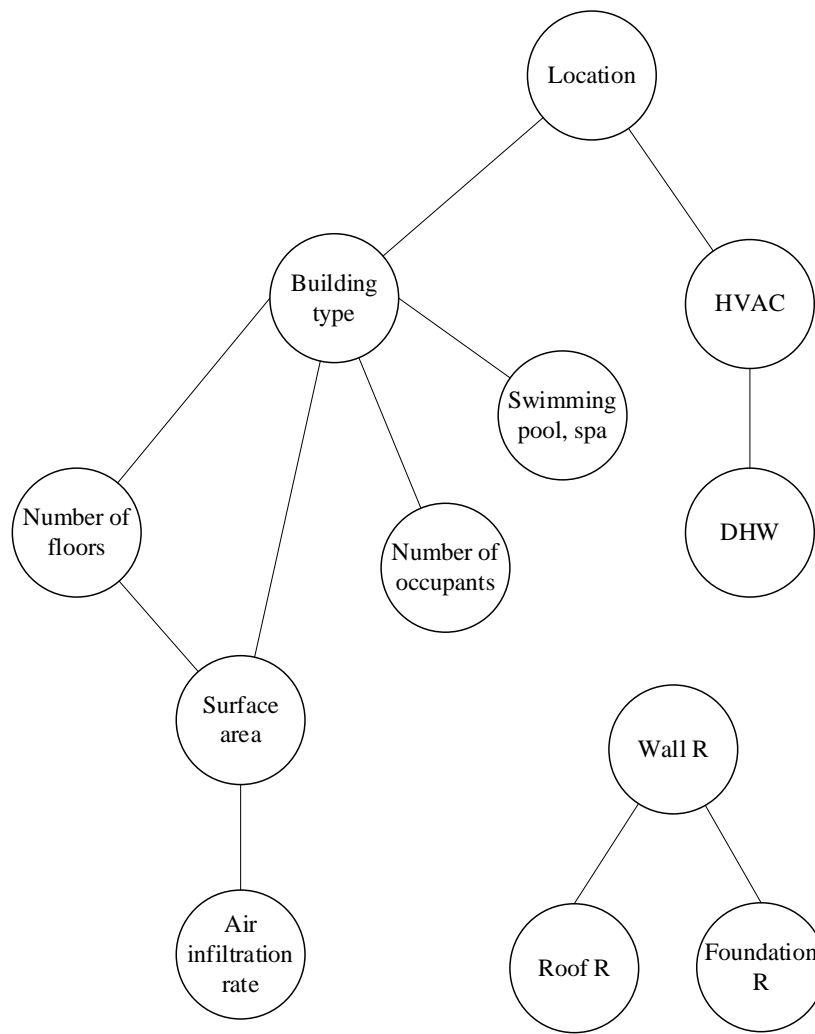


Figure 4.4 Building parameter dependency network. R: thermal resistance of the building envelope, DHW: domestic hot water, HVAC: heating, ventilation and air conditioning

Bayes' theorem, described in Equation (4.8), is applied for each of the connections in the network in Figure 4.4, which allows for the determination of the conditional probability of a parameter given prior evidence.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (4.8)$$

where  $A$  and  $B$  are dependent parameters,  $P(A/B)$  is the conditional probability of  $A$  given  $B$  occurring,  $P(A)$  is the prior probability distribution of  $A$ ,  $P(B)$  is the prior probability distribution

of  $B$ , and  $P(B/A)$  is the prior conditional probability of  $B$  given  $A$  occurring.  $P(B/A)$  is typically based on prior knowledge, i.e. based on data obtained in the literature. Prior probability distributions such as  $P(A)$  can be established based on evidence found in the literature and expressed using Equation (4.9).

$$P(A) = \frac{n_{Ai}}{n} \quad (4.9)$$

where  $n_{Ai}$  is the number of cases for class  $i$  of parameter  $A$  and  $n$  is the number of samples.

As an example of Bayes' Theorem applied to building parameters, the conditional probability distribution for  $P(\text{Roof } R | \text{Wall } R)$ , which is read as *the probability of a building having a certain value of roof thermal resistance given a value of wall thermal resistance*, is illustrated in Table 4.4. The shaded values represent the maximum probability of roof thermal resistance for a given category of wall thermal resistance, i.e. the most likely possibility. Note that “ $R$ ” denotes *thermal resistance*.

Table 4.4 Conditional probability distributions for roof thermal insulation based on wall insulation levels. Shaded values represent maximums

Roof R [m <sup>2</sup> ·K·W <sup>-1</sup> ]	Wall R [m <sup>2</sup> ·K·W <sup>-1</sup> ]				P(Roof R)
	1	2	3	5	
1	0.146	0.041	0.008	0.008	0.043
2	0.319	0.250	0.069	0.068	0.204
3	0.207	0.214	0.109	0.096	0.181
4	0.138	0.194	0.193	0.134	0.186
5	0.095	0.142	0.295	0.190	0.179
8	0.095	0.159	0.326	0.504	0.208

The conditional probability values in Table 4.4 illustrate the dependent nature of the roof insulation based on the level of the wall thermal resistance. For example, based on prior evidence, the probability of a building having a 2.0 m<sup>2</sup>·K·W<sup>-1</sup> roof insulation for a house with 2.0 m<sup>2</sup>·K·W<sup>-1</sup> wall thermal insulation is 0.250. The higher the insulation in the walls, the more probable it is to have higher levels of roof insulation. A non-negligible portion of the building stock has unconventional configurations, which is important to represent correctly in the model. For example, approximately

9.5% of homes with very low wall thermal resistance ( $1.0 \text{ m}^2 \cdot \text{K} \cdot \text{W}^{-1}$ ) have very high roof thermal resistance ( $8.0 \text{ m}^2 \cdot \text{K} \cdot \text{W}^{-1}$ ). This is plausible due to the ease that attic insulation can be retrofitted to higher values of thermal insulation. The opposite is not true however, where homes with high values of wall thermal resistance and low values of roof thermal resistance are rare ( $< 1\%$ ). If the model simply applied the roof insulation prior probability distribution  $P(\text{Roof } R)$ , shown in Table 4.4 in the right-hand column, the virtual buildings produced would not correctly represent the actual state of the building stock. This is consistent with the strategy of developing a virtual residential model that is as close to a building stock model as possible, to produce realistic buildings and improve the classification model development process.

The probability distributions for a number of building parameters are presented in Table 4.5. Each distribution is described in terms of the number of categories, any dependencies based on other characteristics, a histogram depicting the probability mass function, some notes related to that specific category, and any applicable references for the data source.

Table 4.5 Probability distributions for the stochastic parameters generated for each virtual building

Parameter	Categories	Dependencies	Probability distribution	Notes	Reference
Location	1: Rimouski, Québec (L1) 2: Saguenay, Québec (L2) 3: Québec, Québec (L3) 4: Sherbrooke <sup>i</sup> , Québec (L4) 5: Trois-Rivières <sup>i</sup> , Québec (L5) 6: Montréal, Québec (L6) 7: Gatineau, Québec (L7)	N/A	<p>■ QC (Canada)</p>	<sup>i</sup> Also includes several surrounding cities.	Distribution of total number of SFH in the province of Québec (StatCan 2016; NRCan 2011).
Building type	1: Single-detached home (DET) 2: Row house (ROW) 3: Semi-detached home (SDH) 4: Other single-attached <sup>i</sup> (OSA)	Location	<p>■ L1 ■ L2 ■ L3 ■ L4 ■ L5 ■ L6 ■ L7</p>	<sup>i</sup> Other single-attached are residential single-family homes adjacent to non-residential buildings, sharing one or more walls.	Prevalence of each building type by region (StatCan 2016; NRCan 2011).
Number of floors	1: 1 storey 2: 2 storeys	Building type	<p>■ DET ■ ROW ■ SDH ■ OSA</p>	Probability distributions for ROW, SDH and OSA are based on the same data, i.e. non-single-detached homes.	Energuides Housing Database (NRCan 2018).



Parameter	Categories	Dependencies	Probability distribution	Notes	Reference
Occupants	1: 1 occupant 2: 2 occupants 3: 3 occupants 4: 4 occupants 5: 5 occupants <sup>i</sup>	Building type	<p>■ DET ■ SDH ■ ROW ■ OSA</p>	<sup>i</sup> Includes homes with more than 5 occupants	Statistics Canada household data (Statistics Canada 2011; Statistics Canada 2016). Energuide Housing Database (NRCan 2018).
Heated surface area	1: 56-93 [75] <sup>i</sup> m <sup>2</sup> 2: 93-139 [115] m <sup>2</sup> 3: 139-186 [160] m <sup>2</sup> 4: 186-232 [210] m <sup>2</sup> 5: >232 [250] m <sup>2</sup>	Building type, number of floors	<p>■ DET1 ■ DET2 ■ Other1 ■ Other2</p>	<sup>i</sup> Median value. “Other” includes ROW, SDH and OSA building types. “1” or “2” indicate either one- or two-stories.	Energuide Housing Database (NRCan 2018).
Wall thermal resistance <sup>i</sup>	1: 0.5-1.5 [1.0] <sup>ii</sup> m <sup>2</sup> KW <sup>-1</sup> 2: 1.5-2.5 [2.0] m <sup>2</sup> KW <sup>-1</sup> 3: 2.5- 4.5 [3.0] m <sup>2</sup> KW <sup>-1</sup> 5: >4.5 [5.0] m <sup>2</sup> KW <sup>-1</sup>	N/A	<p>■ Wall R</p>	<sup>i</sup> Total wall assembly thermal resistance <sup>ii</sup> Median value in brackets	Energuide Housing Database (NRCan 2018).

Parameter	Categories	Dependencies	Probability distribution	Notes	Reference
Roof thermal resistance <sup>i</sup>	1: 0.5-1.5 [1.0] <sup>ii</sup> m <sup>2</sup> KW <sup>-1</sup> 2: 1.5-2.5 [2.0] m <sup>2</sup> KW <sup>-1</sup> 3: 2.5-3.5 [3.0] m <sup>2</sup> KW <sup>-1</sup> 4: 3.5-4.5 [4.0] m <sup>2</sup> KW <sup>-1</sup> 5: 4.5-5.5 [5.0] m <sup>2</sup> KW <sup>-1</sup> 8: >5.5 [8.0] m <sup>2</sup> KW <sup>-1</sup>	Wall thermal resistance	<p>Wall R: ■ 1.0 ■ 2.0 ■ 3.0 ■ 5.0</p> <p>Roof thermal resistance [m<sup>2</sup>KW<sup>-1</sup>]</p>	<sup>i</sup> Total roof assembly thermal resistance <sup>ii</sup> Median value in brackets	Energuidе Housing Database (NRCan 2018).
Foundation thermal resistance <sup>i</sup>	1: 0.5-1.5 [1.0] <sup>ii</sup> m <sup>2</sup> KW <sup>-1</sup> 2: 1.5-2.5 [2.0] m <sup>2</sup> KW <sup>-1</sup> 3: 2.5-3.5 [3.0] m <sup>2</sup> KW <sup>-1</sup> 4: 3.5-4.5 [4.0] m <sup>2</sup> KW <sup>-1</sup>	Wall thermal resistance	<p>Wall R: ■ 1.0 ■ 2.0 ■ 3.0 ■ 5.0</p> <p>Foundation thermal resistance [m<sup>2</sup>KW<sup>-1</sup>]</p>	<sup>i</sup> Total foundation assembly thermal resistance <sup>ii</sup> Median value in brackets	Energuidе Housing Database (NRCan 2018).
Equivalent Leakage Area	1: 248 cm <sup>2</sup> @4Pa 2: 406 cm <sup>2</sup> @4Pa 3: 556 cm <sup>2</sup> @4Pa 4: 775 cm <sup>2</sup> @4Pa 5: 1426 cm <sup>2</sup> @4Pa	Heated surface area	<p>Bldg. area: ■ 75 ■ 115 ■ 160 ■ 210 ■ 250</p> <p>Equivalent leakage area</p>	N/A	Energuidе Housing Database (NRCan 2018).

Parameter	Categories	Dependencies	Probability distribution	Notes	Reference												
Auxiliary heating	1: Electric 2: Non-electric	Location <sup>i</sup>	<table><caption>Fuel source: Electric, Non-electric</caption><thead><tr><th>Location</th><th>Electric</th><th>Non-electric</th></tr></thead><tbody><tr><td>L6</td><td>0.60</td><td>0.40</td></tr><tr><td>L7</td><td>0.45</td><td>0.55</td></tr><tr><td>Other</td><td>1.00</td><td>0.00</td></tr></tbody></table>	Location	Electric	Non-electric	L6	0.60	0.40	L7	0.45	0.55	Other	1.00	0.00	<sup>i</sup> Other heating sources are more common in regions L6 and L7 of the study	Energuide Housing Database (NRCan 2018).
Location	Electric	Non-electric															
L6	0.60	0.40															
L7	0.45	0.55															
Other	1.00	0.00															
Air conditioning	1: No air conditioning 2: Heat pump <sup>i</sup> 3: Window air conditioner		<table><caption>Air conditioning (A/C)</caption><thead><tr><th>Air conditioning (A/C)</th><th>Probability</th></tr></thead><tbody><tr><td>No A/C</td><td>0.65</td></tr><tr><td>A/C</td><td>0.25</td></tr><tr><td>Window A/C</td><td>0.10</td></tr></tbody></table>	Air conditioning (A/C)	Probability	No A/C	0.65	A/C	0.25	Window A/C	0.10	<sup>i</sup> Air-source heat pumps only.	Energuide Housing Database (NRCan 2018).				
Air conditioning (A/C)	Probability																
No A/C	0.65																
A/C	0.25																
Window A/C	0.10																
Heat pump (heating)	1: No heat pump 2: Heat pump <sup>i</sup> + Auxiliary	Air conditioning <sup>ii</sup>	<table><caption>Heating: No HP, HP + Aux.</caption><thead><tr><th>Air conditioning (A/C)</th><th>No HP</th><th>HP + Aux.</th></tr></thead><tbody><tr><td>No A/C</td><td>1.00</td><td>0.00</td></tr><tr><td>A/C</td><td>0.40</td><td>0.60</td></tr><tr><td>Window A/C</td><td>1.00</td><td>0.00</td></tr></tbody></table>	Air conditioning (A/C)	No HP	HP + Aux.	No A/C	1.00	0.00	A/C	0.40	0.60	Window A/C	1.00	0.00	<sup>i</sup> Air-source heat pumps only. <sup>ii</sup> Homes without A/C or with window A/C rarely had heat pumps for heating.	Energuide Housing Database (NRCan 2018).
Air conditioning (A/C)	No HP	HP + Aux.															
No A/C	1.00	0.00															
A/C	0.40	0.60															
Window A/C	1.00	0.00															

Parameter	Categories	Dependencies	Probability distribution	Notes	Reference
Domestic hot water	1: Electric element 2: Non-electric	Auxiliary heating <sup>i</sup>	<p>DHW: ■ Electric ■ Non-electric</p> <p>Probability</p> <p>Heating system energy source</p>	<sup>i</sup> Homes with non-electric heating are more likely to have non-electric hot water heaters	Energuides Housing Database (NRCan 2018).
Swimming pool	1: Swimming pool 2: No swimming pool	Building type	<p>■ DET ■ SDH ■ ROW ■ OSA</p> <p>Probability</p> <p>Pool</p>	N/A	Real estate database (Realtor.ca 2019). Pool energy calculator (Hydro-Québec 2019a).
Spa	1: Spa 2: No spa	N/A	<p>Probability</p> <p>Spa</p>	Limited information was available on spa distribution by building type.	Spa energy calculator (Hydro-Québec 2019b).

The probability distributions presented in Table 4.5 were developed using a number of sources, most predominantly the Energuide Housing Database (EHD) (NRCan 2018). The EHD consists of over 700,000 homes across Canada that have been audited under a home energy efficiency retrofit program. Probability distributions were established by considering the pre-retrofit characteristics for a subset of roughly 27,000 homes in the Province of Québec. The EHD represented the most detailed information available for many of the building characteristics used in the virtual model.

Cumulative probability distributions were developed from each distribution illustrated in Table 4.5, which are used within the model to assign properties during random number generation.

#### **4.6.4 Fixed parameters**

As with any building simulation, a variety of constant values were used where variable values were not necessary or could not effectively be defined. For example, certain material properties were assigned fixed values consistent with material reference databases, such as for the thermal conductivity of wall insulation or the density of concrete. The impact of using a constant value for such cases was minor compared to other variables in the building simulation. An experienced building simulation modeler would be able to select appropriate values for these types of parameters, and for the purpose of brevity they are not reported here.

#### **4.6.5 Window types**

Establishing the typical window types for the province of Québec requires a specific procedure due to the complexity of the available data. In the Energuide Housing Database (EHD) (NRCan 2018) the windows are described in terms of Hot2000 codes (NRCan 2019b), which are six hexadecimal digits corresponding to the number of glazings, window coating, fill type, spacer type, window type, and the frame material of a window.

Each Hot2000 window code is characterized in terms of the overall U-value and solar heat gain coefficients (SHGC). These values are illustrated in a bubble chart to visualize the distribution of the window properties (Figure 4.5). The k-medoid clustering technique is then applied to the window data set to establish the most representative windows as a function of the number of glazings (single, double and triple). The number of clusters  $k$  is established iteratively to cover at

least 75% of the building stock. Through this method, it is established that 3 single-glazed windows, 15 double-glazed windows, and 6 triple-glazed windows represent the studied building stock. The exact characteristics of the windows are described in Table 4.6.

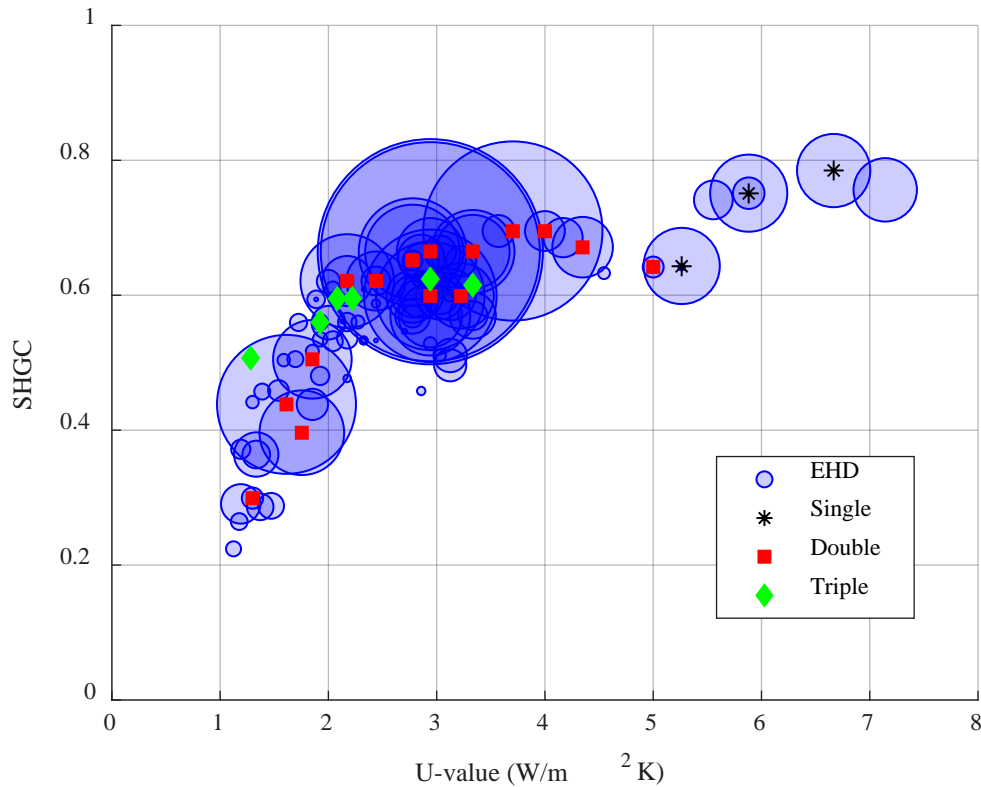


Figure 4.5 Window data by U-value and SHGC with resulting k-medoid clusters for single-, double- and triple-glazed windows. Size of EHD data points represents number of occurrences in the data set for a given window type

The probability distribution for each of the 24 window types is established from the number of occurrences of each window type (Table 4.6). While there are many individual windows reproduced in the model, the objective for window classification is primarily to identify the number of glazings, as opposed to the exact model.

Table 4.6 Window properties and clustering results

Window	Probability	Cumulative probability	Hot2000 code	U-value	SHGC	Glazing	Coating	Fill type	Spacer	Type	Frame
<b>Single-glazed windows</b>											
1 <b>Single 1</b>	0.020	0.020	100000	6.667	0.785	Single	Clear	13mm air	Metal	Picture	Aluminum
2 <b>Single 2</b>	0.026	0.046	100002	5.882	0.751	Single	Clear	13mm air	Metal	Picture	Wood
3 <b>Single 3</b>	0.022	0.068	100022	5.263	0.643	Single	Clear	13mm air	Metal	Slider with sash	Wood
<b>Double-glazed windows</b>											
4 <b>Double 1</b>	0.002	0.070	200040	5.000	0.642	Double	Clear	13mm air	Metal	Patio door	Aluminum
5 <b>Double 2</b>	0.014	0.084	200020	4.348	0.671	Double	Clear	13mm air	Metal	Slider with sash	Aluminum
6 <b>Double 3</b>	0.006	0.090	202000	4.000	0.695	Double	Clear	6 mm air	Metal	Picture	Aluminum
7 <b>Double 4</b>	0.123	0.212	200000	3.704	0.695	Double	Clear	13mm air	Metal	Picture	Aluminum
8 <b>Double 5</b>	0.047	0.259	202002	3.333	0.665	Double	Clear	6 mm air	Metal	Picture	Wood
9 <b>Double 6</b>	0.016	0.276	202012	3.226	0.599	Double	Clear	6 mm air	Metal	Hinged	Wood
10 <b>Double 7</b>	0.394	0.670	200002	2.941	0.665	Double	Clear	13mm air	Metal	Picture	Wood
11 <b>Double 8</b>	0.112	0.782	200012	2.941	0.599	Double	Clear	13mm air	Metal	Hinged	Wood
12 <b>Double 9</b>	0.002	0.784	200006	2.778	0.652	Double	Clear	13mm air	Metal	Picture	Fiberglass
13 <b>Double 10</b>	0.016	0.800	231002	2.439	0.621	Double	Low-e .20 (hard 1)	9 mm air	Metal	Picture	Wood
14 <b>Double 11</b>	0.043	0.843	234002	2.174	0.621	Double	Low-e .20 (hard 1)	9mm argon	Metal	Picture	Wood
15 <b>Double 12</b>	0.024	0.867	223214	1.852	0.505	Double	Low-e .10 (soft)	13mm argon	Insulating	Hinged	Vinyl
16 <b>Double 13</b>	0.027	0.894	213214	1.754	0.396	Double	Low-e .04 (soft)	13mm argon	Insulating	Hinged	Vinyl
17 <b>Double 14</b>	0.074	0.968	213204	1.613	0.438	Double	Low-e .04 (soft)	13mm argon	Insulating	Picture	Vinyl
18 <b>Double 15</b>	0.002	0.969	644204	1.299	0.299	Double - 1 heat mirror	Low-e .35 (hard 2)	9mm argon	Insulating	Picture	Vinyl
<b>Triple-glazed windows</b>											
19 <b>Triple 1</b>	0.002	0.971	300010	3.333	0.615	Triple	Clear	13mm air	Metal	Hinged	Aluminum
20 <b>Triple 2</b>	0.007	0.978	301000	2.941	0.624	Triple	Clear	9 mm air	Metal	Picture	Aluminum
21 <b>Triple 3</b>	0.006	0.984	301002	2.222	0.595	Triple	Clear	9 mm air	Metal	Picture	Wood
22 <b>Triple 4</b>	0.005	0.989	300002	2.083	0.595	Triple	Clear	13mm air	Metal	Picture	Wood
23 <b>Triple 5</b>	0.008	0.997	331002	1.923	0.560	Triple	Low-e .20 (hard 1)	9 mm air	Metal	Picture	Wood
24 <b>Triple 6</b>	0.003	1	323204	1.282	0.507	Triple	Low-e .10 (soft)	13mm argon	Insulating	Picture	Vinyl

### 4.6.6 Climate files

The location of the home determines the climate file used in the building simulation. For the purpose of generating realistic electricity consumption values via building simulation, weather data files are used for the year 2016. Classification based on location will therefore be able to more consistently determine the location of an anonymous smart meter data based on the order of magnitude of the electricity load in a heating-dominated climate. Other climate data can be used to generate the VSM data sets by substituting the regional climate files for other years.

## 4.7 VSM Data

The overall procedure to produce the VSM data set is depicted in Figure 4.2, which shows that the ultimate goal is to produce electricity smart meter data with known building characteristics. A data set of 200,000 VSM profiles is provided with this paper, which consists of input data, VSM data, load profiles, and annual totals for heating, cooling, lighting, equipment and domestic hot water electricity.

### 4.7.1 Input data

A virtual smart meter profile is produced by first generating a single-family home. Each home is characterized by the uniform probability distributions (UPD) and probability mass functions (PMF) mentioned previously. Each distribution requires the generation of a distinct random number, which is then used to determine which value to use for a given parameter. A sample input set is presented in Table 4.7 as an example of the link between the random number generated for each parameter and the corresponding value used in the simulation.

Table 4.7 Sample input set based on random number generation

Parameter	Random number	Bin #	Total # bins	Distribution type	Corresponding value
Location	0.554	6	7	PMF	Montréal, Canada
Building type	0.536	1	4	PMF	Detached
Occupancy profile number	0.350	06 <sup>1</sup>	15	UPD	Profile #6
Window # glazings	0.961	2	3	PMF	Double-glazed windows
Surface area	0.403	3	5	PMF	160 m <sup>2</sup>



Parameter	Random number	Bin #	Total bins	#	Distribution type	Corresponding value
Window-to-wall ratio	0.120	1	3		UPD	0.1
Building rotation	0.162	1	4		UPD	0° rotation
Occupants	0.117	1	5		PMF	1 occupant
Building adjacency	0.561	1	4		UPD	Detached - no adjacency
Floors	0.131	1	2		PMF	1 floor
Wall R <sup>2</sup>	0.441	2	4		PMF	2 m <sup>2</sup> KW <sup>-1</sup>
Roof R	0.110	2	6		PMF	2 m <sup>2</sup> KW <sup>-1</sup>
Foundation R	0.901	2	4		PMF	2 m <sup>2</sup> KW <sup>-1</sup>
Infiltration rate	0.976	5	5		PMF	1426 cm <sup>2</sup>
Air conditioning	0.594	1	3		PMF	No AC
Heat pump	0.272	1	2		PMF	No heat pump
Auxiliary heating type	0.809	1	2		PMF	Electric
DHW type	0.035	1	2		PMF	Electric
Aspect ratio	0.391	2	5		UPD	0.9
Pool	0.242	1	2		PMF	No pool
Spa	0.745	1	2		PMF	No Spa

<sup>1</sup> Requires two digits due to 15 possible bins.

<sup>2</sup> R: Thermal resistance.

The resulting bin numbers from the example above correspond to a unique combination of input values for a particular electricity consumption data profile. The building characteristics can therefore be traced back for each VSM data profile, to the nearest bin value. In some cases the exact value used in the building simulation is also provided, such as for the heated surface area. In the case that a duplicate input set exists, the duplicate is removed in postprocessing and a new profile is generated and added to the data set.

#### 4.7.2 VSM profile data

The total house electricity consumption in kilowatt-hours (kWh) is recorded at 15-minute intervals for 365 days. This accounts for 35,040 data points per VSM profile, not including input data, which

is stored separately. The data set with 200,000 virtual homes requires approximately 50 gigabytes of storage space in an uncompressed format.

### 4.7.3 Load profiles

The internal load profiles generated with the CREST tool are provided with the VSM data for reference. Each profile is based on the number of occupants (from 1 to 5) and a randomly generated profile (from 1 to 15), resulting in 75 different possible load profiles. Lighting, appliance and domestic hot water loads are based on the occupancy profiles. A user can refer to the *Occupancy Profile Number* and *Occupants* inputs to determine the corresponding load profile that was used in the building simulation. Data is organized in terms of the following characteristics:

- Occupants in the home
- Occupants active
- Lighting energy use
- Appliance energy use
- Domestic hot water energy use

### 4.7.4 Annual totals for heating, cooling, lighting, equipment and domestic hot water electricity use

In addition to the smart meter data at 15-minute intervals, annual total electricity use for heating, cooling, lighting, equipment and domestic hot water is also provided. This allows for a greater understanding of the electricity consumption within a home without adding a significant amount of data to an already very large data set. Annual totals are provided in kilowatt-hours (kWh).

### 4.7.5 Overview of the data

The generated homes for the VSM profile framework are intended to cover a range of buildings that represent the most likely combinations of parameters. A box and whisker plot of the annual electricity consumption for the virtual smart meter data set is shown in Figure 4.6, which illustrates

the range of values for heating, cooling, electricity, equipment, domestic hot water (DHW), and total electricity consumption. The box represents the data that lies between the 25<sup>th</sup> and 75<sup>th</sup> percentiles for the data set, i.e. the interquartile range (IQR). The whiskers represent values within 1.5 times the IQR. Data outside the whiskers are considered outliers and marked with individual data points. The median value is indicated by a red line within the box plot. The virtual profiles are compared to the average single-family home electricity consumption for the province of Québec (NRCan 2019).

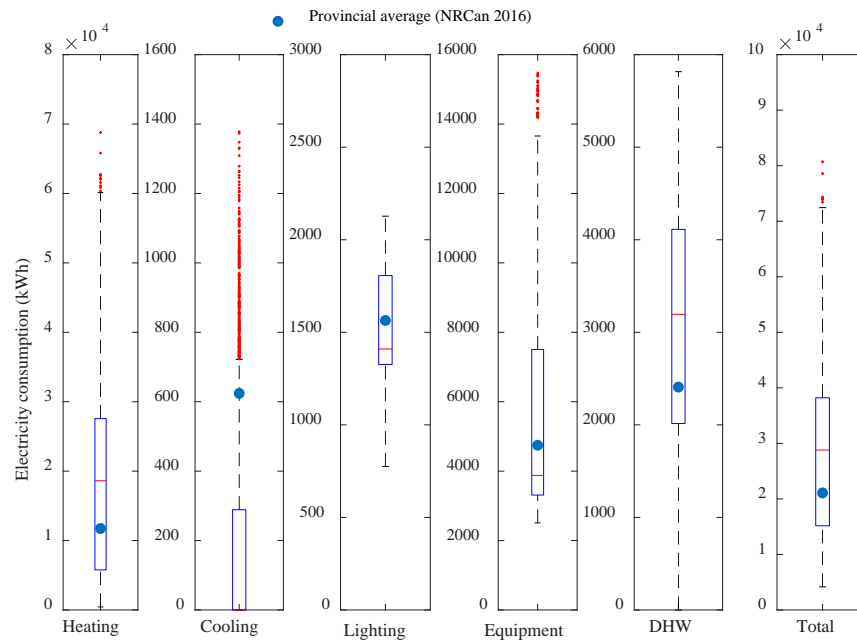


Figure 4.6 Box and whisker plot illustrating the variation in electricity consumption when compared with provincial averages

The virtual homes vary quite widely in terms of electricity consumption, which is consistent with the variety of building parameters used to produce the data set. Buildings with non-electric heating would have negligible heating electricity consumption, while large, poorly insulated homes with large families would have a relatively high overall electricity consumption. Cooling values tend to be underestimated by the model, which is likely due to the underlying assumptions used in the VSM framework that underestimate the number of residences equipped with air conditioning systems in the studied building stock. Equipment values vary widely depending on the presence of

a pool and/or spa in the house. Overall the total electricity consumption is consistent with provincial averages for Québec.

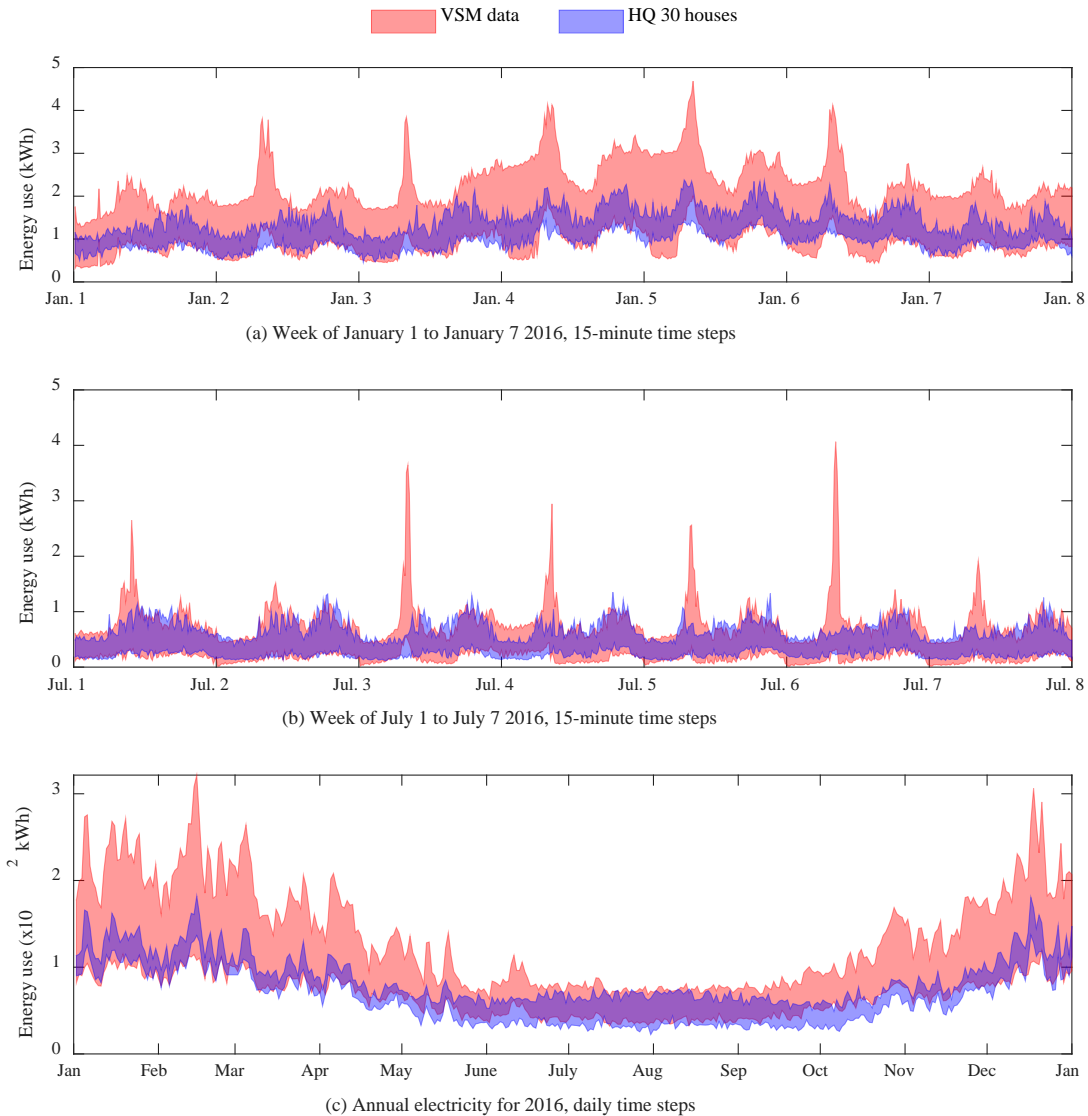


Figure 4.7 Interquartile ranges for the VSM data set and 30 houses in the province of Québec for (a) the first week of January, (b) the first week of July, and (c) daily energy use for a full year.

To further illustrate the range of possible profiles in the virtual smart meter data set, the complete VSM data for all locations is compared to the measured electricity consumption of 30 houses in location #5 (see Table 4.5) in Figure 4.7. Each graph depicts the interquartile range (25<sup>th</sup> to 75<sup>th</sup>

percentiles) for the respective data sets at each time step. The results in Figure 4.7(a) and Figure 4.7(b) illustrate the quarter-hourly electricity consumption for a week in winter and a week in summer, respectively. Daily energy use is compared for a full year in Figure 4.7(c). For all three cases, the measured smart meter data generally falls within the range of values produced in the virtual smart meter data set. Peak energy use tends to be higher in the virtual profiles when compared with the measured data indicating that there are some houses in the VSM data set that are poorly insulated compared to the 30 houses which tend to be better built. Nonetheless, the results in Figure 4.7 indicate that the virtual smart meter data includes a range of cases consistent with the measured data. In the future, a larger measured data set will be used to compare with the VSM data.

## **4.8 Discussion**

The Virtual Smart Meter data set is provided in a format with sufficient accompanying information to be useful for a variety of purposes. By considering the input data, load profiles and corresponding smart meter data, researchers can utilize the data set to verify their own models, study electricity consumption for a variety of housing types, or estimate electricity consumption for a large number of houses, among other possibilities. In this discussion, some commentary on the general methodology for developing a virtual smart meter data set for another building stock is provided. Two example applications for the virtual smart meter data set are then provided: 1) developing inverse models using classification of electricity smart meter data, and 2) verification of load disaggregation algorithms.

### **4.8.1 Developing and using a virtual smart meter data set**

The methodology presented in this paper can be applied towards developing a virtual smart meter data set for a single-family home residential building stock. Suggestions have been made to aid in seeking out sources of information on building and occupant characteristics. The segmentation and characterization processes of building stock modeling are necessary to develop probability distributions for the parameters used in the VSM building generator. These distributions can be used to develop a Bayesian network, such as the one illustrated in Figure 4.4, to produce

combinations of building inputs that actually match the selected building stock. Less common or impossible combinations of building parameters should not be prioritized when producing a virtual smart meter data set. Alternatively, less categories can be used to characterize the building parameters and reduce the overall number of combinations of inputs.

When producing virtual smart meter profiles, parallel processing is an essential component to reducing the overall time required to complete the virtual data set. Cloud computing can allow for multiple building simulations in parallel. In the case of this study, 20 simulations were run in parallel, with each annual simulation requiring approximately 15 seconds. A full data set of 200,000 profiles requires approximately 42 hours to produce under these conditions.

Working with large smart meter data sets can require significant computational resources, whether they are virtual or measured data. The main limitation is in the random access memory (RAM) for loading and processing a significant number of smart meter profiles ( $>100,000$ ), which scales based on the complexity of the building stock and the desired accuracy. This issue can be offset by aggregating the electricity consumption to different time scales, such as hourly or daily data, to significantly reduce the required memory use. Alternatively, studying specific periods of time, such as a single week of data at 15-minute intervals, can reduce the memory requirements to manageable levels. Finally, filtering the smart meter data based on specific building parameters can also be a viable option.

#### **4.8.2 Classification modeling of electricity smart meter data**

As described in the literature review, few public smart meter data sets contain detailed information about the buildings. The data sets that exist have limited information about the houses or are too narrow in the quantity of buildings studied, which restricts the range of parameters studied. As an example, in order to evaluate the impact of the building envelope on the electricity meter data, detailed envelope performance data is required. If a classification modeler wishes to estimate the level of thermal resistance for a home based on the smart meter data, a smart meter data set with known envelope properties is required to train a classification model. The VSM data set can be used for this purpose.

For an example of this method applied in practice, readers can refer to the study by Neale et al. (2019), where linear discriminant analysis was applied to a preliminary version of the VSM data set to predict a number of building parameters for real smart meter data. The classification process significantly increased the accuracy of predicting building parameters when compared to random guessing. Computational resources for classification are also discussed in Neale et al. (2019).

### **4.8.3 Automated load disaggregation algorithms**

The virtual smart meter data set can be used to verify the effectiveness of automated electricity load disaggregation methods, which are commonly used to divide electricity smart meter data into heating, appliance, lighting, and other relevant loads when only aggregated data is available. Such methods are commonly applied to limited data sets, such as those described in Table 4.1, for which there is little variety in the appliance, occupancy and load profiles to test the algorithms. Deb et al. (2019) developed a load disaggregation algorithm for electric heating and tested it on a data set for a single home with 37 days of data. While the house was well-parametrized, the extent of the validation was limited by the scope of the data set.

The VSM data set can be used to test load disaggregation algorithms for a variety of building geometry, occupant behaviour, appliance and lighting configurations and other factors. Researchers can apply their algorithms to the virtual smart meter data and compare it directly to the submetered heating, cooling, lighting, appliance and domestic hot water subtotals. The impact of each building parameter can be studied in order to improve the accuracy of the tested methods.

## **4.9 Conclusion**

A virtual smart meter (VSM) data set is an effective tool for evaluating the impact that building parameters have on electricity consumption. The provided data set and methodology can serve as an example for other researchers on producing and structuring VSM data for other building stocks. In addition, there are many possible applications that require smart meter data with known building characteristics, including load disaggregation algorithms, classification modeling, peak load studies, technology evaluations and other work. The smart meter data sets that currently exist in literature limit the effectiveness of studies in these fields. In addition, the use of a model-based

smart meter data set allows the user to filter the data based on specific inputs to fit the desired approach. The provided VSM data set can be used by researchers to verify their methods and provide insight on residential electricity consumption.

The VSM framework and data sets will be improved in a number of ways in the future. The authors intend to continue to develop the Bayesian network defining the dependencies between the building characteristics. While the VSM framework is not intended to be a building stock model at this stage of the work, eventually it is the hope of the authors to improve the framework to the point where the produced data set is as representative as possible of the building stock, improving upon the results in Figure 4.6. In addition, as additional information is obtained on the building stock the probability distributions for each parameter will be updated. Work is also ongoing to extract building data from smart meter data using classification modeling, which will provide an additional source of information on the building stock.

#### **4.10 Acknowledgements**

The authors gratefully acknowledge the financial support of an IVADO Fundamental Research Grant [PRF-2017-12].

#### **4.11 Data set**

A data set of 200,000 virtual smart meter profiles with corresponding building characteristics can be found at <http://vsmdata.meca.polymtl.ca/>.

#### **4.12 References**

- Akshay Uttama Nambi, S. N., Antonio Reyes Lua, and R. Venkatesha Prasad. 2015. “LocED: Location-Aware Energy Disaggregation Framework.” In BuildSys 2015 - 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments, 45–54. Association for Computing Machinery, Inc. doi:10.1145/2821650.2821659.
- ASHRAE. 2013. ASHRAE Handbook of Fundamentals. Atlanta, GA: ASHRAE.



- Barker, Sean, Aditya Mishra, David Irwin, Emmanuel Cecchet, Prashant Shenoy, and Jeannie Albrecht. 2012. “Smart\*: An Open Data Set and Tools for Enabling Research in Sustainable Homes.” In *Proc. SustKDD '12*, 1–6.
- Beckel, Christian, Leyna Sadamori, Thorsten Staake, and Silvia Santini. 2014. “Revealing Household Characteristics from Smart Meter Data.” *Energy* 78: 397–410.
- Booth, A.T., R. Choudhary, and D.J. Spiegelhalter. 2012. “Handling Uncertainty in Housing Stock Models.” *Building and Environment* 48 (February). Pergamon: 35–47.  
doi:10.1016/J.BUILDENV.2011.08.016.
- Carroll, Paula, Tadhg Murphy, Michael Hanley, Daniel Dempsey, and John Dunne. 2018. “Household Classification Using Smart Meter Data.” *Journal of Official Statistics* 34 (1): 1–25.
- CER. 2012. CER Smart Metering Project - Electricity Customer Behaviour Trial, 2009-2010. 1st ed. Irish Social Science Data Archive. SN: 0012-00.
- Deb, Chirag, Mario Frei, Johannes Hofer, and Arno Schlueter. 2019. “Automated Load Disaggregation for Residences with Electrical Resistance Heating.” *Energy and Buildings* 182 (January). Elsevier Ltd: 61–74. doi:10.1016/j.enbuild.2018.10.011.
- Hammerstrom, D J, R Ambrosio, J Brous, T A Carlon, D P Chassin, J G Desteese, R T Guttromson, et al. 2007. Pacific Northwest GridWise™ Testbed Demonstration Projects; Part I. Olympic Peninsula Project.
- Hammerstrom, D J, J Brous, T A Carlon, D P Chassin, C Eustis, G R Horst, O M Järvegren, et al. 2007. Pacific Northwest GridWise™ Testbed Demonstration Projects; Part II. Grid Friendly™ Appliance Project.
- Hydro-Québec. 2012. Réponse d’Hydro-Québec Distribution Aux Engagements 3, 10, 15, 18, 21 (UC), 21 (UMQ) ET 22 À 27.
- Hydro-Québec. 2016. Rapport Annuel 2015. Montréal, Canada.

Hydro-Québec. 2019a. “Pool Energy Calculator.”

<http://www.hydroquebec.com/residential/customer-space/electricity-use/tools/swimming-pool-calculator.html>.

Hydro-Québec. 2019b. “Spa Energy Calculator.”

<http://www.hydroquebec.com/residential/customer-space/electricity-use/tools/spa-calculator.html>.

IEA (International Energy Agency). 2017. Energy Efficiency 2017.

IEA (International Energy Agency). 2019. “Smart Grids - Tracking Clean Energy Progress.”

<https://www.iea.org/tcep/energyintegration/smartgrids/>.

Johnson, Geoffrey, and Ian Beausoleil-Morrison. 2017. “Electrical-End-Use Data from 23 Houses Sampled Each Minute for Simulating Micro-Generation Systems.” *Applied Thermal Engineering* 114 (March). Pergamon: 1449–1456.

doi:10.1016/J.APPLTHERMALENG.2016.07.133.

Kleiminger, Wilhelm, Christian Beckel, and Silvia Santini. 2015. “Household Occupancy Monitoring Using Electricity Meters.” In 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2015), 975–986.

doi:10.1145/2750858.2807538.

Klein, S.A., W.A. Beckman, J.W. Mitchell, J.A. Duffie, N.A. Duffie, T.L. Freeman, J.C.

Mitchell, et al. 2017. “TRNSYS 18: A Transient System Simulation Program, Solar Energy Laboratory, University of Wisconsin, Madison, USA.” <http://sel.me.wisc.edu/trnsys>.

Makonin, Stephen, Bradley Ellert, Ivan V. Bajić, and Fred Popowich. 2016. “Electricity, Water, and Natural Gas Consumption of a Residential House in Canada from 2012 to 2014.”

*Scientific Data* 3 (160037): 1–12. doi:10.1038/sdata.2016.37.

Mathworks. 2018. “Matlab R2018b.”

McKenna, Eoghan, and Murray Thomson. 2016. “High-Resolution Stochastic Integrated Thermal–electrical Domestic Demand Model.” *Applied Energy* 165: 445–461.

- Murray, David, Lina Stankovic, and Vladimir Stankovic. 2017. “An Electrical Load Measurements Dataset of United Kingdom Households from a Two-Year Longitudinal Study Background & Summary.” *Nature - Scientific Data*. doi:10.1038/sdata.2016.122.
- Neale, Adam, Michaël Kummert, and Michel Bernier. 2018. “Generator : A Stochastic Virtual Smart Meter Data Generation Model for Residential Building Stock Characterization.” In *ESim 2018, the 10th Conference of IBPSA-Canada*, 65–74. Montreal, Canada.
- Neale, Adam, Michaël Kummert, and Michel Bernier. 2019. “Linear Discriminant Analysis for Classification of Building Parameters for a Large Virtual Smart Meter Data Set.” In *Proceedings of the 16th IBPSA Conference*, 8. Rome, Italy.
- NRCan (Natural Resources Canada). 2011. *Survey of Household Energy Use*.
- NRCan (Natural Resources Canada). 2018. “Energuide Housing Database.”
- NRCan (Natural Resources Canada). 2019a. “Residential Sector - Table 1: Secondary Energy Use and GHG Emissions by Energy Source.” *National Energy Use Database*.  
<http://oe.nrcan.gc.ca/corporate/statistics/neud/dpa/showTable.cfm?type=CP&sector=res&juris=qc&rn=1&page=0>.
- NRCan (Natural Resources Canada). 2019b. “Hot2000.” <https://www.nrcan.gc.ca/energy-efficiency/energy-efficiency-homes/professional-opportunities/tools-industry-professionals/20596>.
- NRCan (Natural Resources Canada). 2020. “Canada’s Secondary Energy Use (Final Demand) by Sector, End Use and Subsector.”  
<http://oe.nrcan.gc.ca/corporate/statistics/neud/dpa/showTable.cfm?type=HB&sector=aaa&juris=ca&rn=2&page=0>.
- Realtor.ca. 2019. “Comprehensive Real Estate Listing Database.” <https://www.realtor.ca/>.
- Reinhart, Christoph F., and Carlos Cerezo Davila. 2016. “Urban Building Energy Modeling – A Review of a Nascent Field.” *Building and Environment* 97: 196–202.

- Saldanha, Neil, and Ian Beausoleil-Morrison. 2012. “Measured End-Use Electric Load Profiles for 12 Canadian Houses at High Temporal Resolution.” *Energy and Buildings* 49 (June). Elsevier: 519–530. doi:10.1016/J.ENBUILD.2012.02.050.
- Sherman, M.H., and D.T. Grimsrud. 1980. “Infiltration-Pressurization Correlation: Simplified Physical Modeling.” *ASHRAE Transactions* 86 (2): 778.
- StatCan. 2011. “Household Size, by Province and Territory.” *Census of Population and Statistics Canada Catalogue No. 98-313-XCB*.
- StatCan. 2016. “2016 Census.” Statistics Canada. <http://www12.statcan.gc.ca/census-recensement/2016/dp-pd/index-eng.cfm>.
- Swan, Lukas G., and V. Ismet Ugursal. 2009. “Modeling of End-Use Energy Consumption in the Residential Sector: A Review of Modeling Techniques.” *Renewable and Sustainable Energy Reviews* 13 (8): 1819–1835.

## CHAPTER 5      ARTICLE 2: DISCRIMINANT ANALYSIS

### CLASSIFICATION OF RESIDENTIAL ELECTRICITY SMART METER DATA

Neale, Adam, Michaël Kummert, and Michel Bernier. 2021. *Energy and Buildings*, December. Elsevier, 111823.

#### 5.1 Abstract

The objective of this study is to apply machine learning classification to predict building characteristics from electricity smart meter data for the purpose of building stock characterization. Given that there are no publicly available large-scale residential electric smart meter data sets with detailed building characteristics, an open-source virtual smart meter (VSM) data set is used. The VSM data consists of electricity consumption profiles for 200,000 homes with 21 known characteristics, which are used to train predictive models with linear discriminant analysis (LDA). The classification accuracy (CA) is determined for a variety of scenarios where the meter data aggregation and period are varied. The CA depends on the parameter to be classified (the class), the number of data points per building (the features) and the number of buildings used for classification. Reliable classification results are obtained when the number of buildings exceeds the number of features by a significant margin. An application of the developed predictive models to a small data set of 30 real houses illustrates the usefulness of the method but also the challenges in achieving a generalized model with virtual data.

#### 5.2 Introduction

Evaluating the effectiveness of energy efficiency measures and technology upgrades for buildings on a large scale, such as at the urban, provincial or national levels, can require the use of a building stock energy model. Developing such models can be accomplished using a number of techniques, including top-down models and bottom-up engineering and statistics-based models (Swan and Ugursal 2009). Building archetypes are one such method that requires an information gathering process known as segmentation and characterisation (Sokol et al. 2016). Regardless of the

technique used, information on the building stock is a limiting factor on the accuracy of the resulting model (Booth, Choudhary, and Spiegelhalter 2012).

Electricity smart metering has become very widespread in the last decade, with the United States installing 98 million meters in 2019, the total now covering 70% of the U.S. residential market (Mordor Intelligence 2021). In the province of Québec, Canada, there are 3.7 million installed smart meters, which includes over 1 million single-family homes (Hydro-Québec 2016). The prevalence of metered data can provide a wealth of information considering that the electricity consumption is monitored at subhourly intervals. With a sufficiently large smart meter data set, with information about relevant building parameters, it could be possible to leverage the vast quantity of metered data to extract building details from anonymous smart meter data. Such details could be used for the purpose of building stock modelling techniques, such as building stock segmentation and characterisation (Sokol et al. 2016). The main problem is to leverage such smart meter data in a way that protects the privacy of the homeowners while serving modeling professionals as a source of building stock data.

A variety of well-documented supervised machine learning techniques can serve to establish correlations between a set of inputs and one or more data points. For example, James et al. (2013) provides an introduction and overview of many statistical learning techniques. The term *supervised* indicates that a training set is required to develop a model that correlates a series of input data, or *predictors*, with a corresponding output, or *response*. This process is also often called *classification* when the response is qualitative rather than quantitative, as the process determines a *class* category for a particular set of data, as opposed to a numeric value. A generalized illustration of predictive model development and application is described in Figure 5.1.

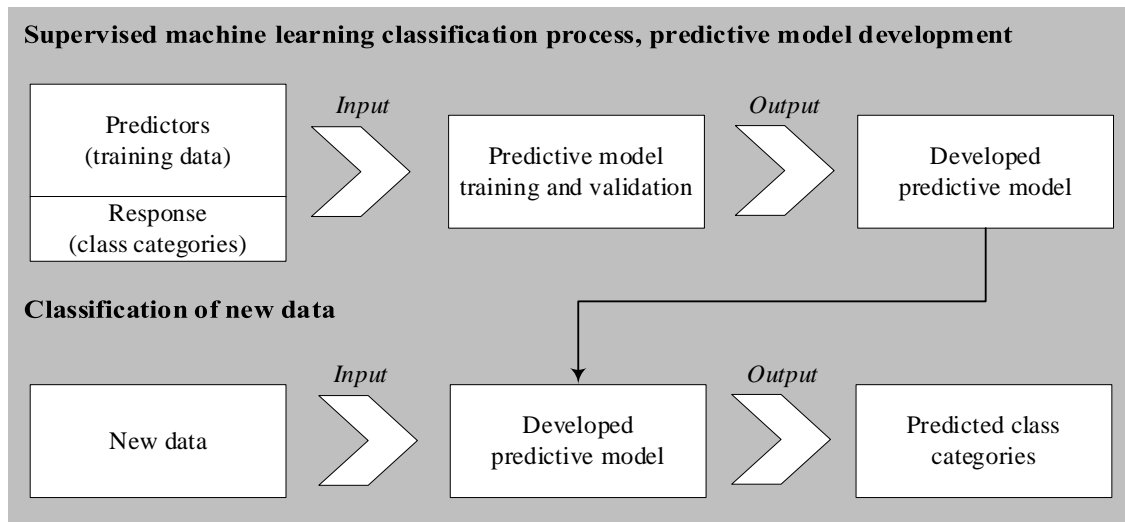


Figure 5.1. Generalized supervised machine learning predictive model development process

The general approach to develop a predictive model involves a training set of predictor data used to explain a qualitative response variable described in terms of class categories, as described in Figure 5.1. The term *predictive model* is used here, as the purpose of classification is to develop a model such that new data can be input to predict a class value (Shmueli 2010). The new predictor data must share the same number of features as the original training data set, which describes the number of data for each case. The classification terms used above, such as predictor, response and feature, are described in more detail in the glossary at the end of the paper.

Statistical learning methods, including the classification approach illustrated in Figure 5.1, are already applied to a wide variety of fields, including handwriting recognition, DNA mapping, e-mail spam detection, etc. (Hastie, Tibshirani, and Friedman 2009). Classification has been identified as one of the key smart grid analysis tools going forward (Y. Zhang, Huang, and Bompard 2018). Supervised machine learning classification could be used on smart meter data provided some information is known about the buildings in the data set, such as the surface area, location, etc. However, studies have shown there are very few residential smart meter data sets with sufficient information about the houses (Neale, Kummert, and Bernier 2020a).

The objective of this study is to leverage a virtual smart meter data set in order to develop predictive models to predict building characteristics from real electricity smart meter data and ultimately

improve the building stock characterization process. Linear discriminant analysis is evaluated as a technique to perform this task. Predictive models are developed for 21 known building parameters using the virtual data set of 200,000 buildings. The influence of the data set size and the feature selection on the classification accuracy is presented. The developed models are applied using a set of 30 houses with real smart meter data with known building parameters to test the generalization of the predictive models. More specifically, this paper aims to:

1. Demonstrate how LDA can be used to classify electricity smart meter data, using a set of virtual smart meter data designed for that purpose;
2. Illustrate the impact of the data set size and number of features on the classification accuracy;
3. Guide future users of LDA on potential problems when developing predictive models on large data sets, with feature selection recommendations for specific building parameters;
4. Demonstrate the limitations of applying real smart meter data to a predictive model developed using virtual data;
5. Present how this approach can be applied for building stock energy model development, a current need in industry.

This paper builds upon a previous work (Neale, Kummert, and Bernier 2019) which laid the groundwork for the present study. This journal paper contains entirely new results, significantly increased detail on the methods, new forms of presentation of the results, additional analysis and conclusions to guide those wishing to use LDA classification.

### **5.3 Literature review**

Data required for housing stock model development is one of the key limiting factors for accurate stock energy prediction (Booth, Choudhary, and Spiegelhalter 2012). Smart meter data presents a potential untapped opportunity for insight into every residential building, but due to privacy reasons it is most often anonymous and without any information on the building's characteristics. A building's parameters could be predicted using classification with a sufficient training set, but to the authors' knowledge no studies have fully evaluated the potential of doing so.



Following are some works describing common smart meter data analytics applications and techniques. The focus of the first portion of the review is to describe common methods and practices related to machine learning in building applications, followed by smart meter data analytics with a few examples. Next, a review targeting previous works in supervised machine learning of smart meter data for the purpose of predicting building parameters is performed. While the focus of the authors is residential energy consumption, where applicable non-residential cases are examined as well.

### 5.3.1 Machine learning in building applications

Machine learning (ML) can be categorized as supervised ML, unsupervised ML, semi-supervised ML and reinforcement learning techniques. Sarker (2021) provides a comprehensive review of machine learning techniques with descriptions and general applications. Supervised learning, which is the focus of this study, can be divided into two categories depending on whether the studied variables are discrete or continuous. Machine learning on discrete variables is referred to as *classification*, while for continuous variables it is known as *regression*. There are many reference texts describing the statistical derivation of methods in machine learning as well, such as the one by Hastie et al. (2009).

Machine learning has become commonplace for a variety of building applications. Djenouri et al. (2019) provide an overview of ML in smart building applications, which summarizes a wide variety of statistical methods that are divided in two broad categories: occupant-centric and energy/device centric applications. Occupant-centric machine learning focuses on occupancy detection, activity recognition and preference/behaviour identification. Energy/device-centric applications include energy profiling and demand estimation, appliance profiling and fault detection, and sensor inference.

The reviews by Sarker (2021) and Djenouri et al. (2019) provide comprehensive descriptions of a variety of ML methods and algorithms. Some specific examples of ML applications in buildings are provided here as well. For example, Gładyszewska-Fiedoruk and Sulewska (2020) applied linear discriminant analysis (LDA) classification and artificial neural networks (ANN) on thermal comfort surveys to evaluate occupant responses to various building indoor environmental

conditions. Esen et al. (2008) use ANN and adaptive neuro-fuzzy inference systems (ANFIS) to forecast the performance of ground-source heat pumps under a variety of conditions. Li et al. (2016) apply LDA to perform fault detection and diagnosis (FDD) on a chiller, which demonstrated the effectiveness of multiscale classification for FDD of mechanical systems in buildings. These studies illustrate how the use of ML has permeated many facets of the field of building engineering, while the focus of the authors is specifically on smart meter data analysis.

### **5.3.2 Smart meter analytics**

Wang et al. (2018) performed a thorough review of smart meter data analytics methods. Applications identified include load analysis, load forecasting, load management and other various subcategories. Techniques include time series analysis, dimension reduction, outlier detection, classification, clustering, deep learning, and more. While Wang identifies classification as a relevant technique, building characterization is not listed in the review. Few cases have been found in the literature of supervised machine learning on smart meter data for residential building characterisation. Many works have used regression and clustering techniques on thermal and electricity metered data, both supervised and unsupervised. Many works are cited by Wang et al. (2018) for interested readers, and a few examples are provided here for context.

Classification and other machine learning (ML) algorithms have been applied to smart meter data in recent works, but in many cases in a context of anomaly detection (L. Zhang et al. 2019; Himeur et al. 2021; Oprea et al. 2021). These works aim to detect unusual meter data that may affect energy analysis techniques, such as load forecasting and/or profiling, as well as energy theft detection. ML has been applied in specific smart meter applications, such as identifying changes in occupant behaviour via metered data pattern recognition, for the purpose of diagnosing at-risk patients in distress (Chalmers et al. 2019). Non-intrusive load monitoring (NILM) is another frequent application where classification and supervised machine learning are applied (Klemenjak 2018).

Gianniou et al. (2018) applied regression techniques to daily thermal energy data to predict temperature setpoint and building envelope characteristics for 14,000 houses in Denmark. While classification was not used in their study, properties of a building stock are successfully extracted from meter data with some degree of accuracy. The study is limited by the information available

on each building, as only the weather, heating energy and basic building geometry were available to develop linear regression models.

Unsupervised support-vector regression was applied by Westermann et al. (2020) on electricity smart meter data to predict heating systems for two sets of 400 buildings. Clustering techniques were applied to identify different energy signatures from metered data to identify heating system types. For a case study applied to British Columbia, Canada, the authors were able to accurately predict the distribution of heating systems corresponding to the provincial average, within 2% per type. However, no actual system data was available to validate the prediction at the building-level.

Ullah et al. (2020) applied deep learning clustering techniques to monthly residential building stock energy data for the purpose of identifying energy consumption patterns. The work identifies clusters of energy consumption levels in stock data, and also analyzes a single house's energy consumption over several years. Self-organizing maps are employed to cluster the data after a detailed encoding process. Very limited information on the buildings is provided by the authors, and the study was primarily effective for illustrating the manner in which buildings in the stock consumed electricity.

### **5.3.3 Classification of smart meter data for building characterization**

As mentioned previously, works in smart meter data analytics for the purpose of building characterization are limited, primarily due to the lack of appropriate smart meter data sets for model training. Specifically, large residential electricity metered data sets accompanied by building parameters such as the surface area, building type, or the number of occupants, are rare.

Recently a large open-source data set of smart meter data with building meta-data has been made available for 1636 non-residential buildings in the U.S. (Miller et al. 2020). Najafi et al. (2021) performed a feature analysis study on the data set using the Random Forests classification algorithm to predict principal building use, performance and operations strategy. The results show the importance of feature selection and the possible classification accuracy that can be obtained by varying the features, but are restricted to non-residential buildings.

The Irish Social Science Data Archive (ISSDA) Commission for Energy Regulation (CER) data set of Irish residential dwellings with over 4000 homes with smart meter data contains mainly demographic data, though building surface area, number of occupants and building type were also included (CER 2012). Beckel et al. (2014) performed a classification study on the CER data set, which showed that linear discriminant analysis could predict the various class categories with classification accuracy between 35% and 80%. The large limitation of the study was that the heating and cooling electricity consumption were not included in the data, as these loads were covered by other energy sources.

Carroll et al. (2018) performed a study using the Neural Networks and Elastic Net Logistics machine learning techniques on smart meter data, again for the purpose of household demographic classification in Ireland. The CER data set was used similarly to Beckel et al. (2014), though with different machine learning techniques. What was particularly interesting about the Carroll et al. (2018) study was how they tested a combination of 21 different feature values representing different aggregated electricity consumption values. While some households with a lower number of occupants could be accurately classified, in general it was difficult to identify the appropriate demographic class category with a high degree of accuracy, at least for the techniques studied.

What the literature reveals is that there are few examples of classification on electricity smart meter data that evaluate the capability to predict a building class category, such as the heated surface area, based only on the electricity consumption values. While many machine learning classification algorithms exist, this study focuses on evaluating linear discriminant analysis (LDA) for a wide variety of cases, to test the robustness and the possible effects of using LDA on very large data sets.

### **5.3.4 Linear discriminant analysis**

Linear discriminant analysis (LDA) is a robust classification algorithm that uses linear projection to establish a decision boundary between two or more data groups. This is accomplished by choosing a projection line perpendicular to the decision boundary that best separates the data groups by maximizing the distance between the means of the data groupings and minimizing the variance of the data sets. A discriminant function  $\delta_c(x)$  can be determined for a set of data for a

class of category  $c$  using Equation (5.1), where the goal is to determine the maximum value of  $\delta_c(x)$ .

$$\delta_c(x) = \mathbf{x}^T \mathbf{K}^{-1} \boldsymbol{\mu}_c - \frac{1}{2} \boldsymbol{\mu}_c^T \mathbf{K}^{-1} \boldsymbol{\mu}_c + \log(p_c) \quad (5.1)$$

where  $\delta_c(x)$  is the discriminant function,  $\mathbf{x}$  contains the classification data,  $\boldsymbol{\mu}_c$  are the mean values of the data set,  $\mathbf{K}^{-1}$  is the inverse of the pooled covariance matrix, and  $p_c$  is the probability of a new data point belonging to class category  $c$ . A more detailed description and derivation of the components of Equation (5.1) is provided in Appendix 5.1.

Equation (5.1) expresses the projection of the mean and covariance of the data sets on a projection axis and establishes a decision line by maximizing the term to the right of the equal sign. Equating the linear projection equations for two class categories ( $\delta_{c_1}(x) = \delta_{c_2}(x)$ ) results in a linear decision boundary that can be used to classify new data points as either class category  $c_1$  or  $c_2$ , for example. An example of LDA applied to energy consumption data with the derivation of the key equations is provided in Appendix 5.1 for interested readers. The example serves to illustrate how LDA can be applied to energy data in the context of predicting information about buildings.

### 5.3.5 Summary and paper organization

As illustrated in the literature review, machine learning has been applied to address a variety of building energy problems. Supervised machine learning classification on electricity smart meter data has been the subject of very few studies, primarily due to the fact that no appropriate data sets exist for predictive model training of building characteristics using metered data. A new study on machine learning classification using linear discriminant analysis on a virtual smart meter data set is therefore presented.

This paper is organized in a number of sections, with the goal to illustrate the supervised machine learning classification of electricity smart meter data. First, the data set used for predictive model training and development is described. Next, the classification results are presented. A general discussion is then provided to elaborate on the outcome of the study and compare the results to previous studies. Finally, some concluding remarks are provided.

## 5.4 Data set description

As described in the literature review, residential data sets for predictive model development are very limited in scope. A virtual smart meter (VSM) data set for residential buildings was developed by Neale et al. (2020a), which consists of 200,000 homes with a variety of known physical characteristics. The data set is available for anyone to download (Neale, Kummert, and Bernier 2020b). The VSM data allows for a variety of parameters and house types to be evaluated.

The virtual buildings and VSM data were generated using building energy simulations with parameters determined based on probability distributions using available building stock details. The 200,000 houses represent a subset that is relatively close to the distribution of approximately 2 million single-family homes across the province of Québec, Canada. Each home is represented by electricity consumption values for a full year at 15-minute intervals, i.e. 35,040 data points per building, and 21 physical characteristics, such as building location, heated surface area, building envelope thermal resistance and more. The building parameter classes and categories are described in Table 5.1 (Neale, Kummert, and Bernier 2020a).

Table 5.1. VSM data class category descriptions (adapted from Neale et al. 2020). PMF: probability mass function, UPD: uniform probability distribution (i.e. no prior knowledge for the building stock). Values in square brackets represent the median value for that category

Class	Categories		Type of distribution	Description
	#	Values		
Location	7	1: Rimouski 2: Saguenay 3: Québec city 4: Sherbrooke 5: Trois-Rivières 6: Montréal 7: Gatineau	PMF	Region in the province of Québec, Canada, where the building is located.
<b>Building dimensions and orientation</b>				
Building type	4	1: Single-detached home (DET) 2: Row house (ROW) 3: Semi-detached home (SDH) 4: Other single-attached (OSA)	PMF	Type of home.
Aspect ratio	5	1: 0.8 2: 0.9 3: 1.0 4: 1.1 5: 1.2	UPD	Aspect ratio of the home, which refers to the ratio of the width (street-facing dimension) to the length.
Surface area (m <sup>2</sup> )	5	1: 56-93 [75] 2: 93-139 [115] 3: 139-186 [160] 4: 186-232 [210] 5: >232 [250]	PMF	Heated surface area bins, from smallest to largest. Note that in addition to the surface area category, the exact surface area of the house within that bracket is also provided in the VSM data set. Values in square brackets represent the mean surface area for that bin.
Window-to-wall ratio	3	1: 0.1 2: 0.15 3: 0.20	UPD	Ratio of window surface area to wall surface area.

Class	Categories		Type of distribution	Description
	#	Values		
Building rotation	4	1: 0° 2: 90° 3: 180° 4: 270°	UPD	Rotation of the building with respect to south (90° increments).
Building adjacency	4	1: No adjacent buildings 2: Eastern wall adjacent 3: Western wall adjacent 4: Both eastern and western walls adjacent	UPD	Configuration of outdoor walls directly adjacent to another building. “Eastern” and “Western” are in reference to the front of the home being south facing.
Floors	2	1: 1-storey 2: 2-storey	PMF	Number of floors in the home.
<b>Building envelope</b>				
Wall thermal resistance (m <sup>2</sup> KW <sup>-1</sup> )	4	1: 0.5-1.5 [1.0] 2: 1.5-2.5 [2.0] 3: 2.5- 4.5 [3.0] 4: >4.5 [5.0]	PMF	Wall thermal resistance value. Values in square brackets represent the mean value for that bin.
Roof thermal resistance (m <sup>2</sup> KW <sup>-1</sup> )	6	1: 0.5-1.5 [1.0] 2: 1.5-2.5 [2.0] 3: 2.5-3.5 [3.0] 4: 3.5-4.5 [4.0] 5: 4.5-5.5 [5.0] 6: >5.5 [8.0]	PMF	Roof thermal resistance value. Values in square brackets represent the mean value for that bin.
Foundation thermal resistance (m <sup>2</sup> KW <sup>-1</sup> )	4	1: 0.5-1.5 [1.0] 2: 1.5-2.5 [2.0] 3: 2.5-3.5 [3.0] 4: 3.5-4.5 [4.0]	PMF	Foundation thermal resistance value. Values in square brackets represent the mean value for that bin.
Overall building thermal resistance (m <sup>2</sup> KW <sup>-1</sup> )	3	1: <1.56 2: 1.56-2.25 3: >2.25	UPD	Derived from the roof, wall and foundation thermal resistance values.



Class	Categories		Type of distribution	Description
	#	Values		
Air leakage area (cm <sup>2</sup> @4Pa)	5	1: 248 2: 406 3: 556 4: 775 5: 1426	PMF	Used to characterize air infiltration.
Window glazings	3	1: Single-glazed windows 2: Double-glazed windows 3: Triple-glazed windows	PMF	Number of window glazings used in the building.
<b>Heating, air conditioning and domestic hot water</b>				
Air conditioning	3	1: No air conditioning 2: Air-source heat pump 3: Window air conditioner	PMF	Air conditioning system used in the building, if any.
Heat pump	2	1: No heat pump 2: Air-source heat pump + Auxiliary	PMF	Heat pump type.
Auxiliary heating type	2	1: Electric 2: Non-electric	PMF	Auxiliary heating system type.
<b>Occupancy information</b>				
Occupancy profile number	15	1-15: Occupant profiles numbered 1 to 15	UPD	Stochastic occupancy load profiles used in the building simulation. 15 distinct stochastic profiles were used.
Number of occupants	5	1: 1 occupant 2: 2 occupants 3: 3 occupants 4: 4 occupants 5: 5 occupants	PMF	Number of occupants in the home.
<b>Other parameters</b>				
DHW type	2	1: Electric 2: Non-electric	PMF	Domestic hot water heater type.
Pool	2	1: Pool 2: No pool	PMF	Pool type.
Spa	2	1: Spa 2: No spa	PMF	Spa type.

Each house from the Neale *et al.* (2020a) VSM data set is accompanied by the corresponding category for the input parameters described in Table 5.1. The house's electricity consumption is therefore paired with a number of different building parameters that can be used for predictive model development.

### 5.4.1 Example data

The electricity consumption of a house depends on the combination of the parameters described in Table 5.1. In order to present classification results, it is relevant to discuss the smart meter data used to train the predictive models. If the smart meter data from the VSM data set is aggregated to annual electricity consumption it can be plotted in terms of the heated surface area category, as illustrated in Figure 5.2.

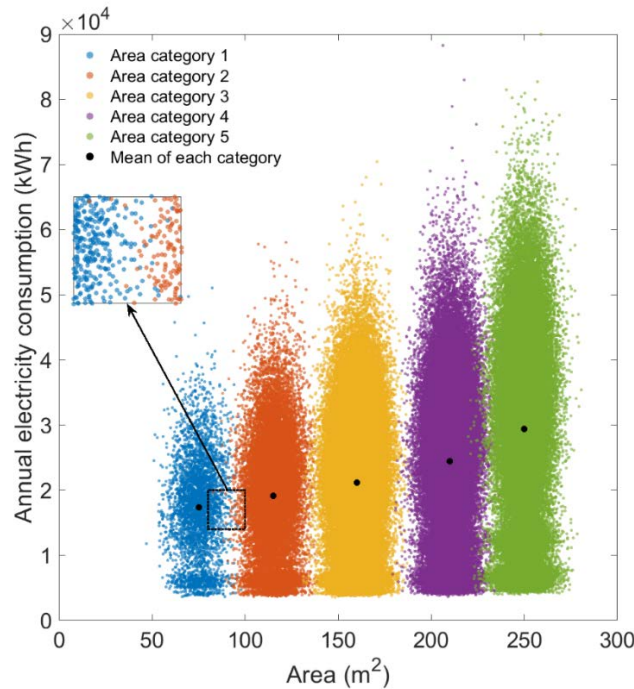


Figure 5.2. Annual electricity consumption of VSM data sorted by surface area category. Each point represents one house with distinct characteristics

The data in Figure 5.2 illustrates how each heated surface area category has a wide range of annual electricity consumption values from the data set. There are gaps between the surface area categories

because the range of values for each category were generated with a Gaussian distribution, making fringe values for a given category less likely, as illustrated in the zoomed-in portion of Figure 5.2. While the categories are somewhat distinct in the figure, it is difficult to determine the size of a house with only the annual electricity consumption. For example, there are houses in all size categories with electricity consumption equal to 20,000 kWh, which indicates that further information is required to accurately classify the houses based on annual electricity consumption alone.

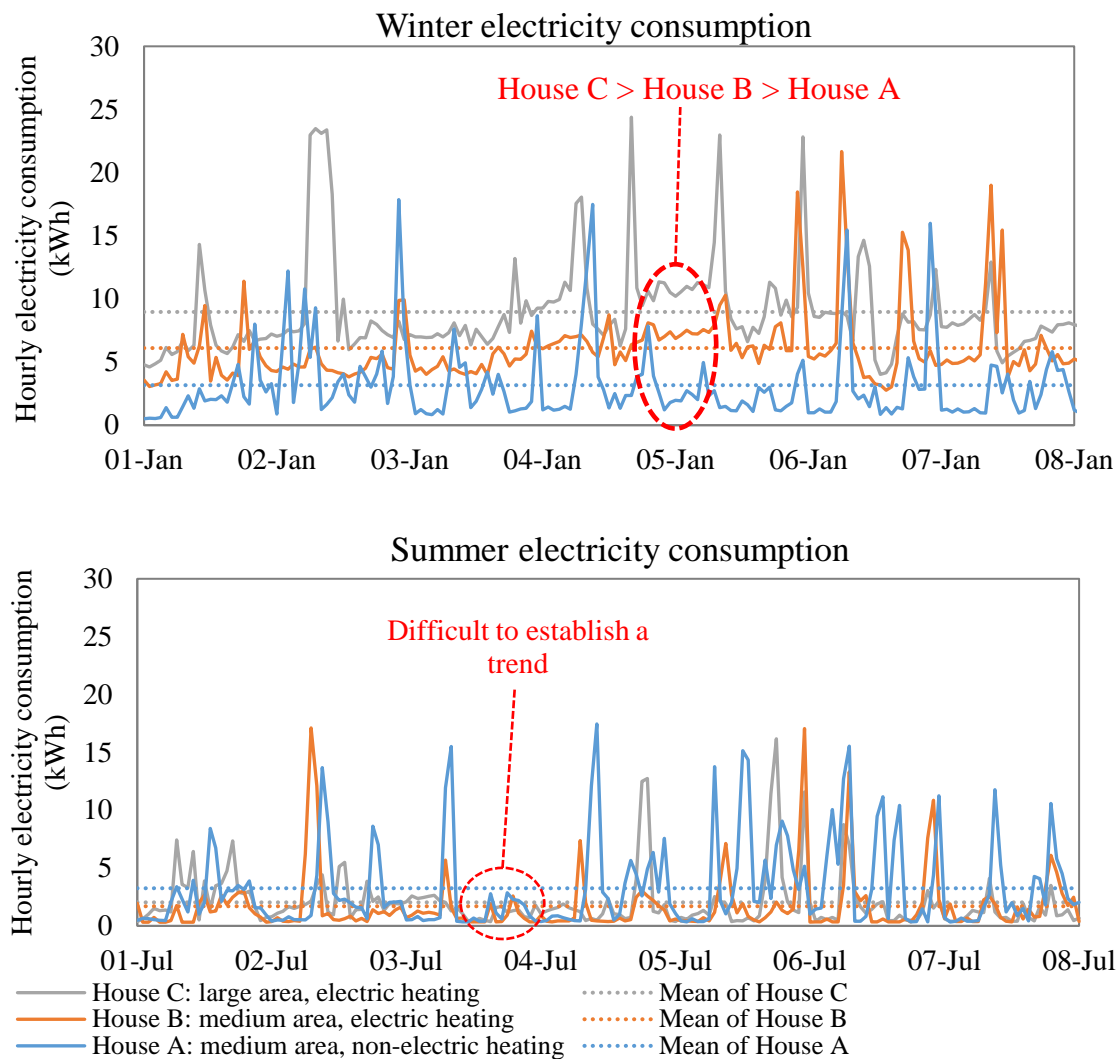


Figure 5.3. Electricity consumption for houses with different characteristics for January (top) and July (bottom) hourly data

The electricity consumption values at hourly aggregation are illustrated in Figure 5.3 for three houses from the VSM data set for a week of data in January and July. A medium house (surface area category 3) without electric heating is shown, which provides an understanding on an electricity consumption profile based primarily on internal loads. Second, a medium house with electric baseboard heating illustrates the variation in electricity consumption when the heating load is included. Finally, a large house (surface area category 5) is compared to the others, with comparatively higher electricity consumption due to the increased overall loads.

The hourly-aggregated data in Figure 5.3 shows many peaks and valleys as the internal loads and the outdoor conditions vary. As indicated in the figure, distinct differences are noticeable in the winter electricity consumption based on the characteristics of the houses, such that visually they can be distinguished based on their size (medium vs. large) or by their heating system type (electric vs. non-electric). These differences are what the classification process aims to identify and associate with the various class categories. The summer electricity consumption illustrates that the profiles for the three illustrated houses are similar and not distinguishable by size or by heating system type, which is logical given the lack of heating load in July for the given building stock.

## 5.5 Classification methodology

As the data in Figure 5.2 illustrates, annual electricity consumption is insufficient for visual classification of homes based on the surface area category, except for extreme cases. A predictive model would need to distinguish the impact of the other characteristics by looking at the electricity smart meter data without knowing those other class categories. For example, a house with 5 occupants has higher variable internal loads than a house with 1 occupant, which should aid in distinguishing between a small house with a large family (high base load, lower heating load) when compared to a large house with a single occupant (low base load, higher heating load).

Linear discriminant analysis as a classification technique for smart meter data is evaluated using a set of 200,000 virtual buildings with a variety of known geometries, internal loads, and heating, ventilation and air-conditioning (HVAC) system parameters. First, a brief description of the methodology behind the predictive model development is provided. The classification accuracy results for a number of scenarios are presented for linear discriminant analysis. Some specific cases

are illustrated in more detail. The time required to produce predictive models based on the number of features is discussed. The impact of the number of buildings is then presented, which illustrates whether 200,000 buildings are required to accurately classify building parameters for the studied residential building stock. The effectiveness of the predictive models developed with the VSM data set are evaluated by applying real smart meter data to the models and comparing the predicted building parameters with known values. Finally, electricity consumption data for a small set of real houses are used to evaluate the prediction capability of the developed models.

### **5.5.1 Predictive model development methodology**

Linear discriminant analysis is evaluated as a method to perform machine learning classification of smart meter data using building characteristics as response variables for electricity consumption predictors. A predictive model is developed using the following general methodology:

1. Select the number of buildings, i.e. 200,000 dwellings.
2. Select the number of features, i.e. hourly annual data has 8760 values and therefore that many features.
3. Build the predictor data matrix from the feature data of each building.
4. Build the response vector from the class categories of each building for the studied class, i.e. the surface area bin for the Area class for each dwelling.
5. Develop the predictive model from the predictor and response data.

The process above is repeated for a variety of configurations. The Matlab Statistics and Machine Learning Toolbox is used for all classification results in this study (Mathworks Inc. 2018).

### **5.5.2 Classification accuracy**

The accuracy of a predictive model is determined based on the number of correct predictions using a validation scheme, as described in Equation (5.13). Predictive models are developed using linear discriminant analysis (LDA) for each building parameter included in the VSM data set. In addition, multiple feature scenarios are presented to study the effect of aggregating the VSM data on the classification accuracy. Scenarios I through IV reflect data for a full year at different time aggregation intervals (monthly, weekly, daily, hourly). Classification for a full year of smart meter

data with subhourly values was not found to be possible due to computer memory limitations and the size of the resulting matrix equation to resolve the classification problem. Scenarios V through VIII are classification results for the month of January with different features (weekly, daily, hourly, and subhourly). Scenarios IX through XII reflect the same feature combinations, but for the month of July. The months of January and July were chosen to evaluate the impact of reducing the data set size and to test the prediction capability of the classification algorithms for building parameters with low or no impact of those parameters during those periods. For example, evaluating the classification of air conditioning using winter data should result in poor classification, since single-family homes typically have zero air conditioning load during the winter months. In addition, it was possible to use subhourly data for the monthly cases, since the number of features was significantly smaller (2976 features) than the annual case (35,040 features). Note that all 200,000 buildings were used for classification, though the impact of the building set size is presented further in the results.

The classification accuracy (CA) results are presented in Table 5.2. Each accuracy value, from 0 to 1, reflects the prediction accuracy for a single predictive model. A value of 0.9 indicates that the class category of 90% of buildings in the data set were correctly predicted with that predictive model. The color scale in the table reflects the range of values within that scale, with red representing close to 0, and green representing close to 1. The best result for each class is indicated with a black border, favoring cases with less features if there is a tie.

The CA results are compared to the accuracy of performing a random guess (RG), which is based on the chance of guessing correctly without knowing any details about the building stock parameters. The RG value is simply  $1/n_{cat}$ , where  $n_{cat}$  is the number of categories in that class, which should be the absolute minimum threshold for classification accuracy. The CA results are above the RG in all cases, indicating the classification algorithm is better than blind guessing. Since the probability distributions for the VSM data set are provided by Neale et al. (2020a), the value of a random guess based on prior knowledge ( $RG_{PK}$ ) can also be determined based on those distributions. This value accounts for the probability of some class categories being more prevalent, which results in a higher chance to guess the correct outcome. Not all parameters had prior

knowledge when the VSM data set was developed and therefore some  $RG_{PK}$  values are not included in Table 5.2.

In order to facilitate the comprehension of the results in Table 5.2 an example is provided. The *Area* class has five categories, which represent five surface area bins described in Table 5.1. The probability of randomly guessing an area bin without any prior knowledge would be 1 in 5, or  $RG = 0.20$ . The *Area* class values were generated using a probability mass function that depended on the type of home (detached, semi-detached, etc.) and the number of floors in the home (Neale, Kummert, and Bernier 2020a). Since those profiles are available, the probability of correctly guessing the class category can be calculated, and in the case of the *Area* class  $RG_{PK} = 0.279$ , slightly better than the blind guess value. The classification accuracy for Scenario I – monthly data for a full year of electricity consumption – is equal to 0.457, which is somewhat better than randomly guessing the category. By increasing the number of features to use hourly data, which corresponds to Scenario IV in Table 5.2, the accuracy improves to 0.793. For this case, the predictive model correctly predicts the surface area category for 4 out of 5 homes in the data set.

To expand upon the *Area* class example, the confusion matrices (CM) for Scenarios I and IV are illustrated in Figure 5.4. The CM provides an understanding of the proportion of correct and incorrect predictions of the predictive model for each category. If a model results in only one category being correctly classified then the model is not very useful when the goal is to identify building characteristics spanning multiple categories.

							TP	TN
True class category	1	2	3	4	5			
	0	0	3827	5	98		<b>0.000</b>	1.000
	0	<b>0</b>	18286	148	1664		<b>0.000</b>	1.000
	0	0	<b>63248</b>	2731	8716		<b>0.847</b>	0.153
	0	0	35625	<b>4620</b>	12994		<b>0.087</b>	0.913
	0	0	21109	3442	<b>23487</b>		<b>0.489</b>	0.511
		1	2	3	4	5		
		Predicted category						

							TP	TN
True class category	1	2	3	4	5			
	<b>2478</b>	755	654	43	0		<b>0.631</b>	0.369
	887	<b>13194</b>	5646	371	0		<b>0.656</b>	0.344
	32	1496	<b>69505</b>	3654	8		<b>0.931</b>	0.069
	1	0	4971	<b>43859</b>	4408		<b>0.824</b>	0.176
	0	0	1674	10632	<b>35732</b>		<b>0.744</b>	0.256
		1	2	3	4	5		
		Predicted category						

Figure 5.4. Confusion matrix for Scenario I (left) and Scenario IV for the Area class categories, labelled 1 through 5. TP: true positive, TN: true negative. Bolded values illustrate the correctly predicted cases

Correct predictions in a confusion matrix are placed in the main diagonal where *Predicted category* = *True category*. The confusion matrix for Scenario I in Figure 5.4 illustrates that categories 1 and 2 were incorrectly assigned to categories 3 to 5. This indicates that with monthly features, the difference in electricity consumption is too subtle for the predictive model to distinguish the smallest house size categories. The True Positive (TP) and True Negative (TN) columns indicate the proportion of correctly and incorrectly predicted buildings in the data set, respectively. In Scenario IV there is a much more even spread of building predictions, and in most cases the results are within one size category of being correctly predicted. The corresponding classification accuracy values can be calculated from the confusion matrices by summing the diagonal values, which are the correct predictions, and dividing by the number of houses in the data set, which in this case is equal to 200,000. For example, the case on the left of Figure 5.4 has 91,355 correct predictions ( $63,428 + 4620 + 23,487$ ) out of 200,000 total houses, and therefore 0.457 classification accuracy.



Table 5.2. Classification accuracy results.  $n_{cat}$ : number of category values for that class, RG: random guess,  $RG_{PK}$ : random guess with prior knowledge. Results with a dark outline indicate the best result for that class

Scenario:	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII			
Period:	1 year	1 year	1 year	1 year	Jan.	Jan.	Jan.	Jan.	July	July	July	July			
Interval:	Monthly	Weekly	Daily	Hourly	Weekly	Daily	Hourly	15-min.	Weekly	Daily	Hourly	15-min.			
Features:	12	52	365	8760	4	31	744	2976	4	31	744	2976	$n_{cat}$	RG	$RG_{PK}$
<b>Location</b>	0.840	0.938	0.970	0.973	0.727	0.929	0.964	0.965	0.598	0.717	0.948	0.956	7	0.143	0.386
<b>Physical properties</b>															
Building type	0.801	0.802	0.859	0.928	0.801	0.799	0.861	0.899	0.801	0.801	0.807	0.812	4	0.250	0.662
Aspect ratio	0.201	0.201	0.201	0.206	0.201	0.201	0.202	0.200	0.201	0.201	0.199	0.200	5	0.200	*
Area	0.457	0.459	0.574	0.793	0.461	0.456	0.649	0.775	0.375	0.416	0.540	0.577	5	0.200	0.279
WWR	0.464	0.516	0.589	0.829	0.368	0.469	0.750	0.782	0.348	0.364	0.525	0.572	3	0.333	*
Rotation	0.259	0.265	0.281	0.286	0.253	0.259	0.290	0.286	0.249	0.250	0.268	0.270	4	0.250	*
Adjacent buildings	0.801	0.802	0.843	0.894	0.801	0.799	0.845	0.872	0.801	0.801	0.807	0.812	4	0.250	0.655
Number of floors	0.682	0.720	0.877	0.940	0.635	0.689	0.907	0.922	0.566	0.606	0.808	0.843	2	0.500	0.510
<b>Building envelope</b>															
Wall thermal resistance	0.590	0.607	0.667	0.838	0.590	0.591	0.755	0.792	0.590	0.590	0.634	0.657	4	0.250	0.432
Roof thermal resistance	0.241	0.266	0.360	0.656	0.238	0.241	0.488	0.552	0.215	0.236	0.301	0.334	6	0.167	0.186
Foundation thermal resistance	0.628	0.698	0.849	0.908	0.538	0.600	0.841	0.853	0.482	0.544	0.803	0.819	4	0.250	0.420
Overall thermal resistance	0.491	0.539	0.632	0.756	0.429	0.464	0.677	0.711	0.363	0.438	0.622	0.643	3	0.333	*
Leakage	0.351	0.418	0.590	0.655	0.332	0.381	0.658	0.642	0.217	0.248	0.436	0.482	5	0.200	*
Window glazings	0.902	0.900	0.897	0.959	0.902	0.902	0.932	0.949	0.902	0.902	0.895	0.900	3	0.333	0.818
<b>HVAC</b>															
Air conditioning	0.777	0.817	0.902	0.981	0.739	0.772	0.819	0.817	0.698	0.763	0.926	0.953	3	0.333	0.509
Heat pump	0.952	0.948	0.988	1.000	0.910	0.946	0.997	0.999	0.838	0.843	0.923	0.924	2	0.500	0.729
Auxiliary heating	0.986	0.993	1.000	1.000	0.987	0.984	1.000	1.000	0.731	0.799	0.959	0.959	2	0.500	0.607
<b>Occupancy details</b>															
Occupants	0.506	0.799	1.000	1.000	0.370	0.607	1.000	1.000	0.416	0.748	1.000	1.000	5	0.200	0.246
Profile number	0.270	0.892	1.000	1.000	0.139	0.622	1.000	1.000	0.252	0.808	1.000	1.000	15	0.067	*
<b>Other parameters</b>															
DHW type	0.960	0.963	1.000	1.000	0.957	0.958	1.000	1.000	0.772	0.817	1.000	1.000	2	0.500	0.649
Pool	1.000	1.000	1.000	1.000	0.835	0.835	0.835	0.833	0.992	0.994	1.000	1.000	2	0.500	0.725
Spa	0.919	0.950	1.000	1.000	0.900	0.900	1.000	1.000	0.903	0.904	0.910	0.894	2	0.500	0.820

\* No prior knowledge for the probability distribution for the category values existed for this parameter, therefore no improvement can be made over simply randomly guessing the outcome.

The results in Table 5.2 illustrate a wide range of classification accuracy values that depend on the scenario and the class. Some classes are not well classified, such as the building rotation or aspect ratio. LDA does not perform any better than randomly guessing for these cases, which indicates that these parameters in the VSM data set have little influence on the smart meter data. Other classes have significantly better accuracy, especially as the number of features increases. The number of occupants and the occupancy activity level is easily classified with daily and hourly data, which is likely due to the programmed nature of the profiles and is one of the limitations of simulation-based occupancy models.

In order to visualize the impact of the number of features on the classification accuracy, the data in Table 5.2 can be expressed in graphical form. The CA results for the location, area, air infiltration rate and overall envelope thermal resistance are illustrated in Figure 5.5.

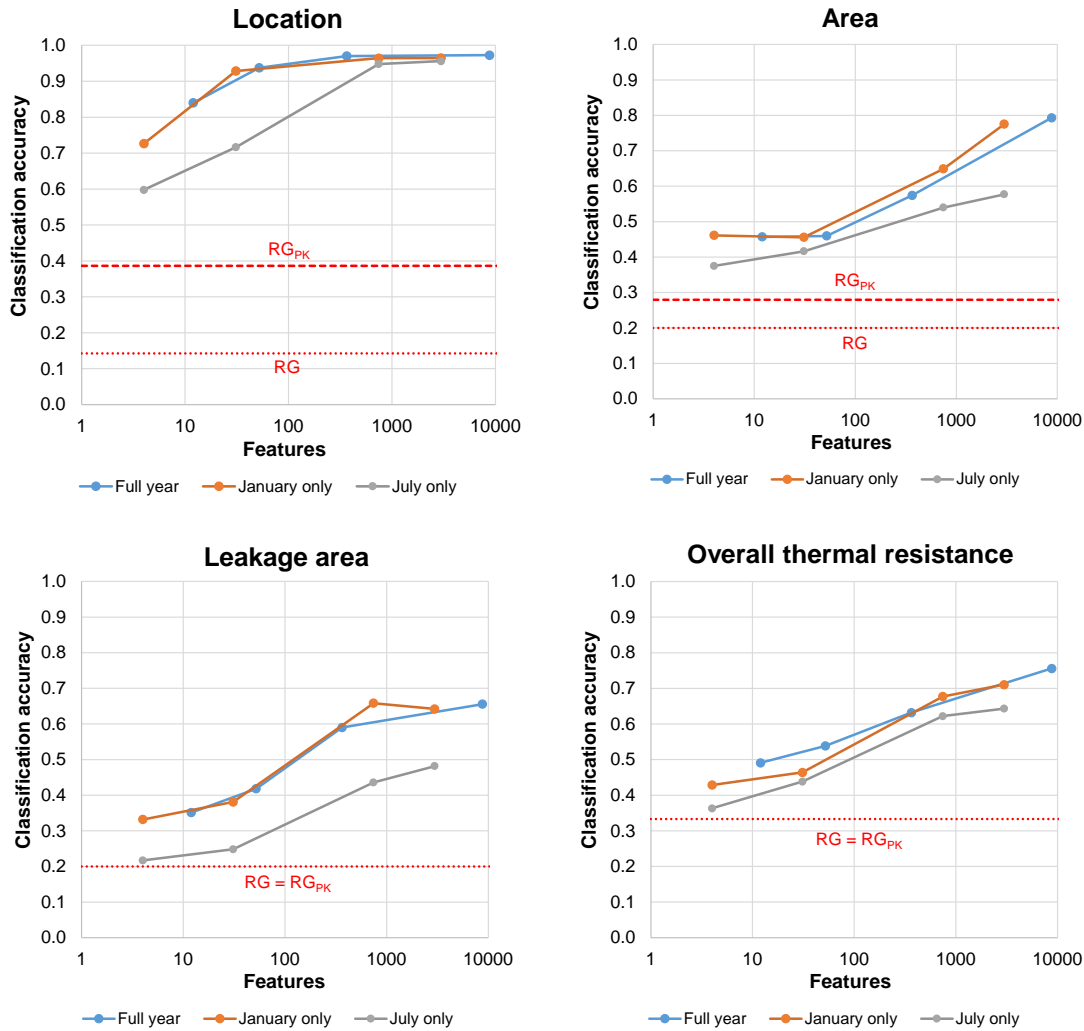


Figure 5.5. Classification accuracy per feature for the location, heated surface area, air infiltration and overall thermal resistance parameters. RG: random guess,  $RG_{PK}$ : random guess based on prior knowledge

The accuracy for LDA predictive models typically increases with a higher number of features, indicating in many cases additional granularity in the electricity consumption is beneficial for classification for building parameters. In some cases, the accuracy reaches a plateau at higher feature values, indicating that there is little gain for increasing the complexity of the predictive

model, such as for the Location parameter. For other parameters, such as the Area, increasing the data granularity further could be beneficial, but at significant computational cost.

Finally, the classification accuracy for January and July smart meter data is illustrated in Figure 5.6 for all cases where  $abs(CA_{jan} - CA_{july}) > 0.005$ , i.e. if there is a meaningful difference between the January and July classification results with the same number of features. Other cases where  $-0.005 \leq (CA_{jan} - CA_{july}) \leq 0.005$  are set equal to  $\pm 0.005$  for visibility but are considered to have a negligible difference for practical purposes. Figure 5.6 illustrates whether there is a difference between summer and winter classification by qualitatively comparing the class results. Each bar in Figure 5.6 represents one pair of CA results that demonstrate the improvement in accuracy using summer or winter data for classification. Results on the negative x-axis illustrate cases where the July data resulted in better classification, while results on the positive x-axis demonstrate cases where January data offered better outcomes. The bar length is a relative difference and thus does not illustrate the absolute accuracy, though these can be obtained in Table 5.2.

The results in Figure 5.6 demonstrate a logical link between the class and the preferred data set to use for classification, when choosing between available winter or summer data. Parameters that impact the heat gain and losses in a home are better classified using January data, as the larger temperature difference between indoor and outdoor leads to proportionately larger heat losses, and therefore more easily detectable differences between the class categories. Similarly, systems related to the heating load of a house are better classified with winter data. Domestic hot water loads, which are tied to ground water temperature, are also better classified in winter.

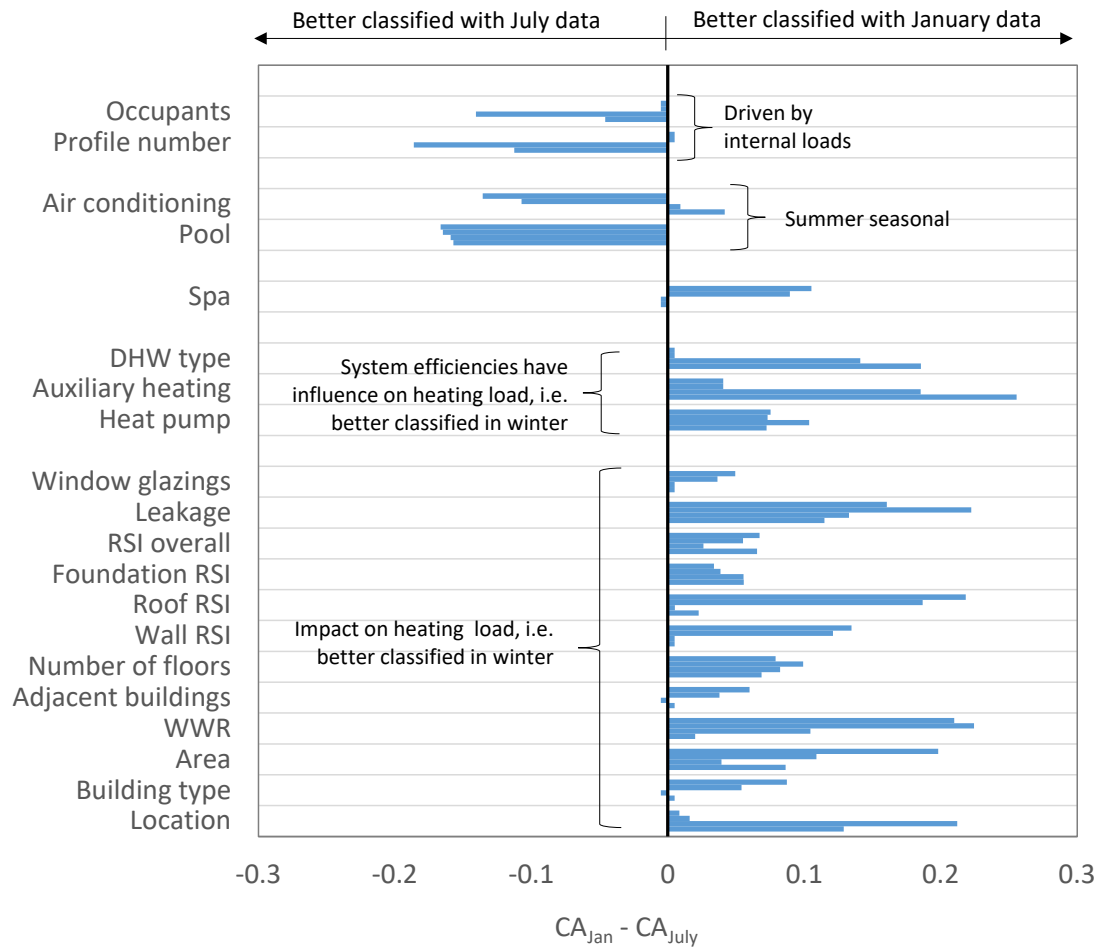


Figure 5.6. Comparison of January and July classification accuracy

Some classes are better classified with summer data, such as the seasonal air conditioning and pool loads. Occupancy-driven internal loads are easier to detect in summer periods due to the lower or non-existent effect of the cooling loads during these periods. The classification process appears to have an easier task at differentiating different occupancy patterns and number of residents using July data. The results in Figure 5.6 illustrate that it is worth considering what parameter is being classified when choosing between seasonal smart meter data sets.

### 5.5.3 LDA predictive model development time

Increasing the number of features in a classification problem will exponentially increase the size of the matrix equation required to solve for the classification boundaries, resulting in increased computation time. The time to compute each predictive model result in Table 5.2 for each feature scenario was recorded. The average time per feature across all classes is illustrated in Figure 5.7. Classes with a higher number of categories tend to take longer as there are additional classification boundaries between each category. All predictive model development was performed on an Intel Core i9-7920X processor @2.9 GHz, 128 GB of RAM @2133 MHz and a SATA III solid-state hard drive.

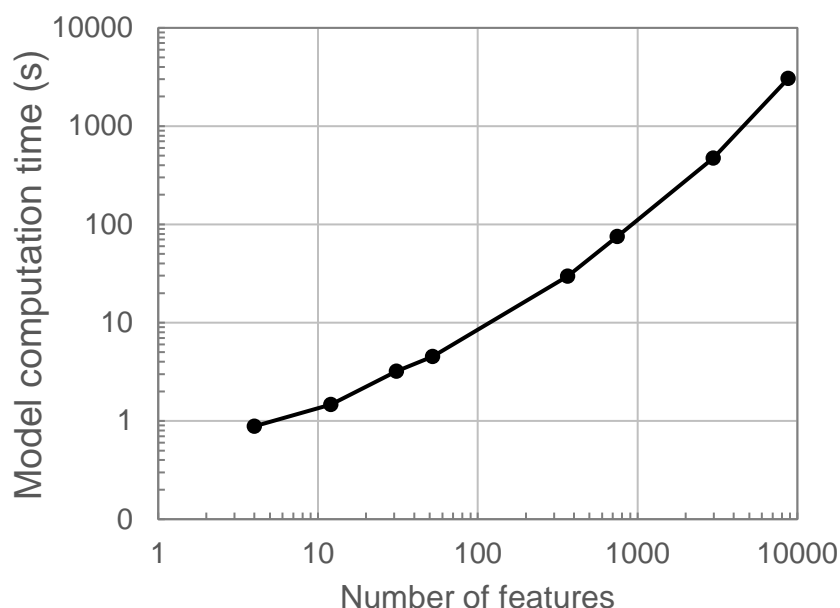


Figure 5.7. Average model computation time based on the number of features

As the number of features increases, the time to compute the predictive model increases exponentially. For an annual predictive model with hourly data, the average computation time is approximately 50 minutes. It should be noted that using 5-fold cross-validation significantly increases the overall time, as it repeats the predictive model process for each fold, using 80% of the data for training and 20% for validation. In addition, the parallel processing features of the

Matlab Parallel Computing Toolbox are used to expedite the calculation process for the presented results (Mathworks Inc. 2018). The chief limitation of the predictive model development is not the time to resolve the model but the quantity of random access memory (RAM) required to store and process the data, with 128 GB RAM being insufficient for some cases.

#### **5.5.4 Impact of data set size on classification accuracy**

The ratio of the data set size to the number of features influences the classification accuracy (Hua et al. 2005). In the case of the present study, the number of buildings in the VSM data set determines the size and the aggregation interval determines the number of features. By testing various subsets of buildings for monthly, weekly, daily and hourly electricity consumption, the range of classification accuracy values is illustrated in Figure 5.8. The y-axis describes the classification accuracy and the x-axis, which is on a logarithmic scale, describes the number of buildings used to develop the predictive models for classification. The range of values represented by the shaded area illustrates the effect of developing a predictive model with different sets of buildings. The smaller the amount of buildings, the more likely the chance of a statistically unrepresentative sample, which results in highly variable classification accuracy.

For example, Figure 5.8(a) shows that testing various sets of 10 buildings using monthly data for the Area class resulted in classification accuracy values ranging from 0.00 to 0.78 (i.e. the range of values plotted on the y-axis). This is due to the effect of the very small sample of buildings and high variability in characteristics of those buildings. Conversely, sets of 5000 buildings with daily data (notation in Figure 5.8(c)) resulted in a much smaller range of values for the Area class, from 0.486 to 0.516. The values at the extreme right of each curve in Figure 5.8 correspond to those in Table 5.2, for Scenarios I through IV for the Area class. For the latter example of Figure 5.8(c), which corresponds to Scenario III (1 year of daily features) for Area, the classification accuracy is equal to 0.574.

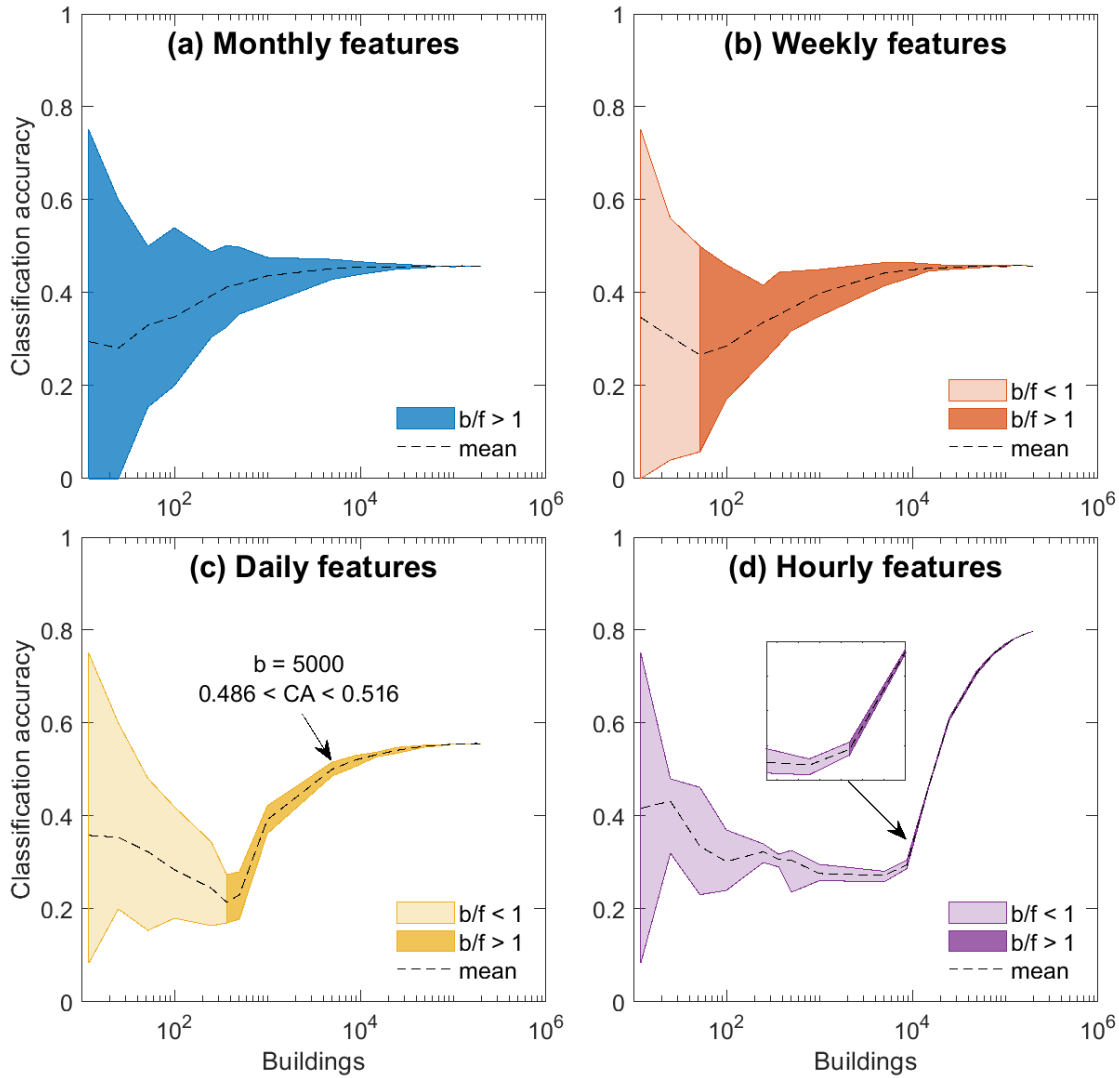


Figure 5.8. Area classification accuracy by building data set size for (a) monthly (12), (b) weekly (52), (c) daily (365) and (d) hourly (8760) features.  $b$ : number of buildings in data set,  $f$ : number of features, CA: classification accuracy

The large range in classification accuracies for a smaller number of buildings can be explained by the likelihood of obtaining the correct predictions by chance. As the number of buildings increases and the characteristics diversify in the data set, the predictive model development stabilizes. This transition occurs when the number of buildings ( $b$ ) is approximately equal to the number of features ( $f$ ), or at  $b/f \cong 1$ , which is represented in the figure as the transition point between the lighter and



darker shaded areas in each graph. The monthly case has only one shaded region as the classification process requires more than 12 buildings to be effective, and therefore  $b/f > 1$  for all developed predictive models for this case. As the number of buildings used in the data set further increases, the range of classification accuracy values narrows and the mean accuracy steadily increases, as depicted in Figure 5.8.

For smart meter data, it is therefore important to have a sufficiently sized sample of buildings to train the model if reliable classification results are desired. With hourly data, this indicates at least 8760 buildings with a variety of characteristics are required. Regardless of the data aggregation scheme used, the classification accuracy stabilized with additional buildings and increased for the daily and hourly feature cases. The curves illustrating the mean values of the accuracy demonstrate the increased performance based on the number of buildings.

### 5.5.5 Application of the developed predictive models on real smart meter data

The predictive model development process is illustrated in Figure 5.1, which describes how a developed model can be used to predict category values for new data. Using real smart meter data as inputs to a model developed with virtual data has some limitations. As an example, a virtual data set based on building energy simulations require occupancy models that inevitably differ from real occupants. These differences can result in unreliable classification, as the underlying assumptions in the virtual model can never perfectly match the reality. However, as mentioned in the literature review, there are currently no appropriate data sets based on real buildings to explore classification of building characteristics, and thus virtual smart meter data sets are the best option for now.

In order to present some of the limitations inherent to classification and guide future research in supervised machine learning of smart meter data, the predictive models developed and presented in Table 5.2 are used to predict building parameters for 30 houses with measured smart meter data, subsequently referred to as *real smart meter (RSM)* data. This general approach is described in Figure 5.1, where new predictor data is input to a predictive model in order to determine the class category for that data. In this case, the RSM data is input to the developed models using the VSM data set, which provides an evaluation of the generalization of the predictive models (Shmueli 2010).

On average the RSM data was missing 3% of electricity consumption data, which were filled using recommended metered data processing techniques (Fowler et al. 2015). For some building characteristics, the true class values for the houses in the RSM data set are available as well, which are used to compare to the predicted categories. The data for the VSM and RSM sets were both for the calendar year 2016. The characteristics of the 30 RSM houses are presented in Table 5.3. All houses are located in *Location 5: Trois-Rivières* (Table 5.1).

Table 5.3. RSM house characteristics. AC: air conditioning. Cat: class category according to Table 5.1

House	Area m <sup>2</sup>	Cat	Occupants	Air conditioning	Pool	Spa	Window glazings
1	191.4	4	4	No AC	Pool	No spa	Double
2	191.4	4	2	Heat pump	No pool	No spa	Double
3	198.2	4	4	No AC	Pool	No spa	Double
4	180.6	3	5	Heat pump	No pool	No spa	Double
5	106.7	2	2	No AC	No pool	No spa	Double
6	145.7	3	4	Heat pump	No pool	No spa	Double
7	162.8	3	1	No AC	No pool	Spa	Double
8	162.1	3	2	No AC	Pool	No spa	Double
9	214.2	4	5	No AC	Pool	Spa	Double
10	204.9	4	4	No AC	Pool	No spa	Double
11	229.5	4	5	Heat pump	No pool	Spa	Double
12	181.8	3	4	No AC	Pool	No spa	Double
13	334.5	5	6*	No AC	Pool	No spa	Double
14	258.7	5	4	No AC	No pool	No spa	Double
15	185.8	3	4	Heat pump	No pool	Spa	Double
16	139.4	3	2	Heat pump	No pool	Spa	Triple
17	204.0	4	5	No AC	Pool	No spa	Double
18	188.0	4	1	Heat pump	No pool	No spa	Double
19	152.2	3	2	No AC	No pool	Spa	Double
20	268.4	5	4	No AC	No pool	No spa	Double
21	179.8	3	2	Heat pump	Pool	No spa	Double
22	152.9	3	3	No AC	No pool	Spa	Double
23	151.1	3	3	Heat pump	No pool	Spa	Double
24	170.0	3	2	No AC	No pool	Spa	Double
25	188.8	4	3	Heat pump	Pool	No spa	Double
26	167.2	3	6*	Heat pump	Pool	No spa	Double
27	346.3	5	2	Heat pump	Pool	Spa	Double
28	330.3	5	4	No AC	No pool	No spa	Double
29	144.0	3	1	No AC	No pool	No spa	Double
30	188.1	4	4	Heat pump	No pool	No spa	Double

\* The VSM data set only contained data for up to 5 occupants, therefore houses with 6 occupants are considered to have 5 instead.

The classification approach for the seven known RSM house parameters is described in Table 5.4. Class categories are assigned to each house based on the VSM parameters in Table 5.3.

Classification accuracy is determined based on the similarity of the predicted category when compared with the true category, with exact matches described as “correct predictions” and with similar matches described as “close predictions”. The definition of “close” varies by parameter and is included in the results to illustrate when classification obtains outcomes equal to or near the correct prediction. For example, if the number of occupants is predicted one higher or lower than a house’s true occupancy, this is considered a close prediction. Some parameters, such as whether a pool is installed in the home, have no “close” option, as they are either correct or incorrect.

Table 5.4. House data set known parameters and definitions for a correct and close prediction

Parameter	Correct prediction	Close prediction
<b>Location</b>	Correct location predicted	Predicted as the correct location or another location with similar heating degree-days (HDD). ±130 HDD
<b>Area</b>	Correct area category predicted	Predicted as the correct area category or one size category larger or smaller. ±40 m <sup>2</sup>
<b>Occupants</b>	Correct number of occupants predicted	Predicted as the correct number of occupants or one occupant more or less than the correct number. ±1 occupant
<b>Air conditioning</b>	Correct AC type predicted	Predicted as the correct air AC type, or if a heat pump predicted as a window air conditioner, or vice versa.
<b>Pool</b>	Presence of a pool correctly predicted	Not applicable
<b>Spa</b>	Presence of a spa correctly predicted	Not applicable
<b>Window glazings</b>	Correct number of window glazings predicted	Not applicable

The classification accuracy (CA) is determined as described in Equation (5.13), which is based on the correct predictions (CP) divided by the total predictions (TP). If applicable, close predictions are substituted in Equation (5.13) for the correct predictions. Classification accuracy for the RSM data is denoted as  $CA_{RSM}$ , which are presented in Table 5.5.  $CA_{RSM}$  results are compared to the accuracy of randomly guessing the categories of each class based on the prior knowledge of the building stock from the VSM data set. The  $RG_{PK}$  is used as a reference since the RSM houses are part of the same building stock as the VSM data.

The results in Table 5.5 illustrate the classification accuracy when the real smart meter data is input to the predictive models developed with the VSM data and the predicted class category is compared to the real class category. As an example, an accuracy of 0.433 indicates that 13 out of 30 houses in the RSM data set had the category correctly predicted. This value can be directly compared to the  $RG_{PK}$  column to evaluate the performance of LDA when compared to a random prediction. If  $CA_{RSM} > RG_{PK}$ , the classification algorithm represents an improvement over a random guess.

Classification of the real smart meter data with linear discriminant analysis has variable accuracy depending on the class, the number of features, and the period used for the smart meter data. There is at least one scenario for each class that resulted in a better prediction than randomly guessing. The average CA improvement for the best scenario for each class is equal to 0.187 and ranges from 0.078 to 0.355. Scenarios with less features generally performed better, which indicates that aggregating the electricity consumption improves the classification accuracy. This is likely due to the way internal loads were generated in the VSM data set used to train the predictive models. Since it is unlikely to match occupant behavior to real data at a subhourly or hourly frequency, aggregating those data for classification seems to be the more reliable approach.

Table 5.5. Classification accuracy results for the real smart meter data set. Best classification results have bold text and borders

Class	Prediction type	Scenario ( <i>period-aggregation-features</i> )												RG <sub>PK</sub>
		1 year-M-12	1 year-W-52	1 year-D-365	1 year-H-8760	Jan-W-4	Jan-D-31	Jan-H-744	Jan-SH-2976	July-W-4	July-D-31	July-H-744	July-SH-2976	
Location	Correct	<b>0.433</b>	<b>0.433</b>	0.267	0.267	0.000	0.167	0.233	0.167	0.000	0.067	0.100	0.033	0.078
	Close	0.667	0.667	0.633	<b>0.700</b>	0.167	0.567	0.633	0.567	0.233	0.233	0.533	0.400	0.282
Area	Correct	0.467	0.233	0.167	0.167	<b>0.500</b>	<b>0.500</b>	0.133	0.100	0.467	0.400	0.133	0.133	0.300
	Close	0.833	0.533	0.433	0.500	<b>0.867</b>	<b>0.867</b>	0.367	0.433	0.833	0.767	0.400	0.433	0.433
Occupant	Correct	0.300	0.167	0.133	0.333	0.267	0.100	0.267	0.167	<b>0.400</b>	0.300	0.267	0.200	0.216
	Close	0.667	<b>0.700</b>	0.667	0.567	0.433	0.400	0.533	0.567	0.567	0.600	0.667	0.667	0.312
Air conditioning	Correct	0.333	0.300	0.500	0.533	0.500	<b>0.567</b>	0.367	0.333	0.533	0.533	0.500	0.433	0.488
	Close	0.433	0.433	0.500	<b>0.700</b>	0.500	0.567	0.400	0.467	0.533	0.533	0.533	0.467	0.033
Pool	Correct	0.600	0.600	0.600	0.600	0.600	0.600	0.500	0.600	<b>0.800</b>	0.667	0.567	0.467	0.563
Spa	Correct	0.500	0.400	0.533	0.500	0.667	0.633	0.400	<b>0.733</b>	0.600	0.533	0.467	0.467	0.650
Window glazings	Correct	<b>0.967</b>	0.900	0.600	0.100	<b>0.967</b>	<b>0.967</b>	0.267	0.167	<b>0.967</b>	<b>0.967</b>	0.100	0.100	0.873

In summary, linear discriminant analysis had mixed results predicting the class categories for a number of building characteristics for a real small data set. The combination of period and aggregation of the electricity consumption that resulted in the best classification result varied by parameter, which further supports the need for additional studies in classification of smart meter data. The data set of real houses used in the present study was quite limited in the number of houses available and the amount of known parameters that could be used for validation purposes. In addition, a non-negligible fraction of data was missing for the real houses, which certainly affects the classification prediction. The impact of the missing data is supported by the fact that aggregating the electricity data often resulted in better predictions. Nevertheless, LDA did demonstrate an improvement over random guessing for all parameters, at least for specific data scenarios. A larger, more detailed RSM data set would provide a better understanding of the link between the classification accuracy, number of features, data set size and number of buildings in the data set.

## 5.6 General discussion

The literature review illustrated the prevalence of machine learning in building applications and the lack of previous studies in classification of buildings based on smart meter data. The study of Beckel et al. (2014) compared multiple classification techniques to predict building characteristics using smart meter data, though most classes were related to occupancy. LDA predicted the floor area category ( $<100 \text{ m}^2$ ,  $100 \text{ to } 200 \text{ m}^2$  or  $>200 \text{ m}^2$ ) for homes in the Beckel study with an accuracy of 45%, compared with up to 80% in this study. For building type (detached or attached), Beckel's study classified houses with 60% accuracy, compared to 93% in this study. The number of occupants was predicted with approximately 70% accuracy, compared to 100% in this study. Carroll et al. (2018) performed a similar analysis as Beckel for occupancy classification, averaging 61% accuracy with different classification algorithms.

This study improves upon previous classification works by systematically analysing the impact of a significant number of scenarios on the classification problem, which guides future classification modelers on the correct way of approaching smart meter data classification. The open-source VSM data set used to train the predictive models represents a new source of data for classification problems that has yet to be fully explored. Until a real smart meter data set with a variety of

measured and surveyed building characteristics is released, the virtual data represents the best data set for smart meter classification studies.

While this paper provides a detailed evaluation of linear discriminant analysis using the VSM data, further work evaluating other classification algorithms using additional metrics would guide those seeking to perform supervised machine learning classification. Other algorithms may improve the results for smart meter data classification. When developing a virtual smart meter data set for classification, some parameters may not be worth attributing distinct class categories, such as differentiating between heat pumps and window air conditioners with similar coefficient of performance (COP) values. Care must be given when attributing class categories and when modeling specific physical behavior, such as the properties of windows installed in a house, as these can only be identified by classification if they were included in the original data set. In addition, a detailed monitoring campaign of real houses with surveyed building characteristics would greatly assist in the validation process of classification studies.

## 5.7 Conclusion

Building stock energy modeling requires a significant amount of information to accurately represent the wide range of building types. This paper seeks to illustrate how supervised machine learning classification with linear discriminant analysis (LDA) can accurately predict building parameters from electricity smart meter data. The virtual smart meter (VSM) data developed by Neale et al. (2020a) is a residential smart meter data set with detailed information on building characteristics, such as building surface area, thermal resistance of the building envelope, occupants, air leakage rate, etc. The VSM data was developed with classification in mind, and the present study uses LDA to evaluate the effectiveness of classification to predict building parameters based solely on electricity smart meter data. Data periods and aggregation intervals are varied to test a number of different feature combinations for each class.

Linear discriminant analysis can effectively classify electricity smart meter data, with classification accuracy values that depend on the parameter studied and the number of features. The building data set size has an important influence on the reliability of the classification outcome. At the very minimum, it is essential to have at least as many buildings ( $b$ ) as the number of features ( $f$ ) in the

data set ( $b/f > 1$ ). As this ratio increases, the classification accuracy for LDA tends to reach an asymptotic value when  $b \gg f$ . This indicates that for a building data set with highly variable characteristics and for parameters better classified with hourly or subhourly features, many buildings are required to develop a reliable predictive model.

Classification accuracy is related to the impact of a building parameter on the electricity consumption. Parameters such as building rotation and aspect ratio are not well classified by LDA. This could be related to the way they are implemented in the building simulation environment used to create the VSM data set. Nevertheless, LDA performs no better than randomly guessing for these two particular parameters.

Other parameters had significantly better classification accuracy than randomly guessing and in some cases reached 100% accuracy, i.e. all 200,000 buildings had their class category accurately predicted by the predictive model. Classification accuracy is strongly tied to the number of features used to develop the model, with higher numbers of features generally resulting in higher accuracy. However, increasing the feature count significantly slows the predictive model development time and increases the memory requirements, as the equations required to resolve the classification problem scale exponentially. There is therefore a significant compromise between accuracy, computation time and computational resources.

One example of a parameter with high classification accuracy is the Occupants parameter. The VSM data set has 15 different profiles for 1 to 5 possible occupants, resulting in 75 different occupancy profiles. While this appears to be a high number of different cases, the classification algorithm can easily detect the differences between each number of occupants and between each profile. Practically speaking the profiles themselves are unlikely to correspond exactly to real house occupants, and so a developed predictive model based on occupancy simulations must be applied with caution.

Applying smart meter data from 30 houses to the developed predictive models resulted in variable classification accuracy. Classification was more effective for aggregate electricity consumption, leading to the conclusion that the stochastic loads of the virtual data set did not fully correspond with the real house occupants. The authors recommend using aggregated electricity consumption



to more easily correspond between modeled occupancy and real occupancy, should the need arise. A more detailed and more extensive real smart meter data set would allow for a better validation of the predictive models developed using the virtual set. Unfortunately, to the knowledge of the authors an appropriate data set for residential building classification does not exist, especially given the conclusions of this paper on the number of buildings required for reliable classification with higher numbers of features.

The results of this paper illustrate that classification has the potential to aid in the segmentation and characterisation of residential building stocks, provided a sufficiently detailed smart meter data set exists to train the models. This would directly benefit those seeking to develop building stock energy models but lack information about the buildings in the studied stock. Given the highly stochastic nature of residential electricity consumption, which depends on the individuals inhabiting the house, proper classification of real smart data using a virtual set of data is not guaranteed. It would be preferable to use a sufficiently large, detailed real smart meter data set with knowledge of the building characteristics to train the predictive model. The virtual classification results illustrates that building parameters can be predicted with a high level of accuracy with the electricity consumption only, which is a promising outcome as a source of data for future building stock energy modeling work.

## 5.8 Glossary

**Features (f):** *Features* are the number of data points representing a single building's energy consumption. By default a single building is represented by 35 040 features, which are electricity consumption data at 15 minute intervals for a full year. Other values are possible since the energy data can be aggregated, for example 365 features for daily-aggregated electricity consumption, or 12 features for monthly-aggregated data.

**Predictors (p):** The *predictors* represent the complete energy consumption data used to develop the predictive models. The predictor data set is a  $[f \times b]$  matrix, where  $f$  is the number of features and  $b$  is the number of buildings, for example  $[365 \times 1000]$  for a data set of daily energy use values of 1000 buildings.

**Class (c) and category (cat):** The *class* is the building parameter selected for classification, such as the building heated surface area or the location of the building. Each class is divided into a number of *categories*, which typically represent bins of values, or discrete values, and do not represent real numbers. For example, locations 1 through 7 represent different regions in the selected building stock, or the air conditioning (AC) class categories may be represented by 1 (no air conditioning), 2 (air-source heat pump), or 3 (window air conditioning). In the latter example, a building's AC would be represented by a value from the set  $\{1,2,3\}$ . The exact values of the categories depend on the chosen data set used for classification.

**Response (r):** The *response* data is the set of class values for one specific parameter, such as the building's surface area category or the location. Each building is represented by a single known value resulting in a vector of length  $b$  with values corresponding to the *class categories* for the parameter studied. Using the example from the *class categories* for air conditioning, the vector  $[r_{AC}]$  of length  $b$  would contain air conditioning class values from the set  $\{1,2,3\}$ . These values are used to train the predictive model by establishing a link between the predictors (energy data) and responses (building parameters).

## 5.9 References

- Beckel, Christian, Leyna Sadamori, Thorsten Staake, and Silvia Santini. 2014. “Revealing Household Characteristics from Smart Meter Data.” *Energy* 78: 397–410.
- Booth, A.T., R. Choudhary, and D.J. Spiegelhalter. 2012. “Handling Uncertainty in Housing Stock Models.” *Building and Environment* 48 (February). Pergamon: 35–47.  
doi:10.1016/J.BUILDENV.2011.08.016.
- Carroll, Paula, Tadhg Murphy, Michael Hanley, Daniel Dempsey, and John Dunne. 2018. “Household Classification Using Smart Meter Data.” *Journal of Official Statistics* 34 (1): 1–25.
- CER. 2012. CER Smart Metering Project - Electricity Customer Behaviour Trial, 2009-2010. 1st ed. Irish Social Science Data Archive. SN: 0012-00.
- Chalmers, Carl, William Hurst, Michael Mackay, and Paul Fergus. 2019. “Identifying Behavioural Changes for Health Monitoring Applications Using the Advanced Metering Infrastructure.” *Behaviour & Information Technology* 38 (11). Taylor and Francis Ltd.: 1154–1166. doi:10.1080/0144929X.2019.1574900.
- Djenouri, Djamel, Roufaida Laidi, Youcef Djenouri, and Ilanko Balasingham. 2019. “Machine Learning for Smart Building Applications.” *ACM Computing Surveys (CSUR)* 52 (2). ACM PUB27 New York, NY, USA . doi:10.1145/3311950.
- Esen, Hikmet, Mustafa Inalli, Abdulkadir Sengur, and Mehmet Esen. 2008. “Artificial Neural Networks and Adaptive Neuro-Fuzzy Assessments for Ground-Coupled Heat Pump System.” *Energy and Buildings* 40 (6). Elsevier: 1074–1083.  
doi:10.1016/J.ENBUILD.2007.10.002.
- Fowler, K.M., A.H. Colotelo, J.L. Downs, K.D. Ham, J.W. Henderson, S.A. Montgmoery, S.A. Parker, and C.R. Vernon. 2015. *Simplified Processing Method for Meter Data Analysis*. Oak Ridge, TN.

- Gianniou, Panagiota, Christoph Reinhart, David Hsu, Alfred Heller, and Carsten Rode. 2018. "Estimation of Temperature Setpoints and Heat Transfer Coefficients among Residential Buildings in Denmark Based on Smart Meter Data." *Building and Environment* 139 (July). Pergamon: 125–133. doi:10.1016/J.BUILDENV.2018.05.016.
- Gładyszewska-Fiedoruk, Katarzyna, and Maria Jolanta Sulewska. 2020. "Thermal Comfort Evaluation Using Linear Discriminant Analysis (LDA) and Artificial Neural Networks (ANNs)." *Energies* 2020, Vol. 13, Page 538 13 (3). Multidisciplinary Digital Publishing Institute: 538. doi:10.3390/EN13030538.
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. 12th print. Springer-Verlag.
- Himeur, Yassine, Khalida Ghanem, Abdullah Alsalemi, Faycal Bensaali, and Abbes Amira. 2021. "Artificial Intelligence Based Anomaly Detection of Energy Consumption in Buildings: A Review, Current Trends and New Perspectives." *Applied Energy*. Elsevier Ltd. doi:10.1016/j.apenergy.2021.116601.
- Hua, J., Z. Xiong, J. Lowey, E. Suh, and E. R. Dougherty. 2005. "Optimal Number of Features as a Function of Sample Size for Various Classification Rules." *Bioinformatics* 21 (8). Oxford Academic: 1509–1515. doi:10.1093/bioinformatics/bti171.
- Hydro-Québec. 2016. *Rapport Annuel 2015*. Montréal, Canada.
- James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning, with Applications in R*. Edited by Gareth M. James. 1st editio. Springer. doi:10.1007/978-1-4614-7138-7.
- Klemenjak, Christoph. 2018. "On Performance Evaluation and Machine Learning Approaches in Non-Intrusive Load Monitoring." *Energy Informatics* 1 (S1). Springer Science and Business Media LLC: 36. doi:10.1186/s42162-018-0051-1.
- Li, Dan, Guoqiang Hu, and Costas J. Spanos. 2016. "A Data-Driven Strategy for Detection and Diagnosis of Building Chiller Faults Using Linear Discriminant Analysis." *Energy and Buildings* 128 (September). Elsevier: 519–529. doi:10.1016/J.ENBUILD.2016.07.014.

- Mathworks Inc. 2018. “Matlab Statistics and Machine Learning Toolbox R2018b.” Natick, Massachusetts, United States.
- Miller, Clayton, Anjukan Kathirgamanathan, Bianca Picchetti, Pandarasamy Arjunan, June Young Park, Zoltan Nagy, Paul Raftery, Brodie W. Hobson, Zixiao Shi, and Forrest Meggers. 2020. “The Building Data Genome Project 2, Energy Meter Data from the ASHRAE Great Energy Predictor III Competition.” *Scientific Data* 7 (1). Nature Research: 1–13. doi:10.1038/s41597-020-00712-x.
- Mordor Intelligence. 2021. “Global Smart Meters Market | Growth, Trends, Forecasts (2020 - 2025).” <https://www.mordorintelligence.com/industry-reports/global-smart-meters-market-industry>.
- Najafi, Behzad, Monica Depalo, Fabio Rinaldi, and Reza Arghandeh. 2021. “Building Characterization through Smart Meter Data Analytics: Determination of the Most Influential Temporal and Importance-in-Prediction Based Features.” *Energy and Buildings* 234 (March). Elsevier Ltd: 110671. doi:10.1016/j.enbuild.2020.110671.
- Neale, Adam, Michaël Kummert, and Michel Bernier. 2019. “Linear Discriminant Analysis for Classification of Building Parameters for a Large Virtual Smart Meter Data Set.” In *Proceedings of the 16th IBPSA Conference*, 3393–3400. Rome, Italy.
- Neale, Adam, Michaël Kummert, and Michel Bernier. 2020a. “Development of a Stochastic Virtual Smart Meter Data Set for a Residential Building Stock – Methodology and Sample Data.” *Journal of Building Performance Simulation* 13 (5): 583–605. doi:10.1080/19401493.2020.1800096.
- Neale, Adam, Michaël Kummert, and Michel Bernier. 2020b. “Virtual Smart Meter Data Set.” <https://vsmdata.meca.polymtl.ca/>.
- Oprea, Simona-Vasilica, Adela Bâra, Florina Camelia Puican, and Ioan Cosmin Radu. 2021. “Anomaly Detection with Machine Learning Algorithms and Big Data in Electricity Consumption.” *Sustainability* 2021, Vol. 13, Page 10963 13 (19). Multidisciplinary Digital Publishing Institute: 10963. doi:10.3390/SU131910963.

- Sarker, Iqbal H. 2021. "Machine Learning: Algorithms, Real-World Applications and Research Directions." *SN Computer Science* 2: 160. doi:10.1007/s42979-021-00592-x.
- Shmueli, Galit. 2010. "To Explain or to Predict?" *Statistical Science* 25 (3): 289–310. doi:10.1214/10-STS330.
- Sokol, Julia, Carlos Cerezo Davila, Christoph F. Reinhart, C. Cerezo, and Christoph F. Reinhart. 2016. "Validation of a Bayesian-Based Method for Defining Residential Archetypes in Urban Building Energy Models." *Energy and Buildings* 134. Elsevier B.V.: 11–24.
- Swan, Lukas G., and V. Ismet Ugursal. 2009. "Modeling of End-Use Energy Consumption in the Residential Sector: A Review of Modeling Techniques." *Renewable and Sustainable Energy Reviews*. doi:10.1016/j.rser.2008.09.033.
- Ullah, Amin, Kilichbek Haydarov, Ijaz Ul Haq, Khan Muhammad, Seungmin Rho, Miyoung Lee, and Sung Wook Baik. 2020. "Deep Learning Assisted Buildings Energy Consumption Profiling Using Smart Meter Data." *Sensors* 2020, Vol. 20, Page 873 20 (3). Multidisciplinary Digital Publishing Institute: 873. doi:10.3390/S20030873.
- Wang, Yi, Qixin Chen, Tao Hong, and Chongqing Kang. 2018. "Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges." *IEEE Trans. Smart Grid*, 24. doi:10.1109/TSG.2018.2805.
- Westermann, Paul, Chirag Deb, Arno Schlueter, and Ralph Evins. 2020. "Unsupervised Learning of Energy Signatures to Identify the Heating System and Building Type Using Smart Meter Data." *Applied Energy* 264 (April). Elsevier: 14. doi:10.1016/J.APENERGY.2020.114715.
- Zhang, Leping, Lu Wan, Yong Xiao, Shuangquan Li, and Chengpeng Zhu. 2019. "Anomaly Detection Method of Smart Meters Data Based on GMM-LDA Clustering Feature Learning and PSO Support Vector Machine." In *ISPEC 2019 - 2019 IEEE Sustainable Power and Energy Conference: Grid Modernization for Energy Revolution, Proceedings*, 2407–2412. Institute of Electrical and Electronics Engineers Inc. doi:10.1109/iSPEC48194.2019.8974989.

Zhang, Yang, Tao Huang, and Ettore Francesco Bompard. 2018. “Big Data Analytics in Smart Grids: A Review.” *Energy Informatics* 1 (1). Springer Science and Business Media LLC: 8. doi:10.1186/s42162-018-0007-5.

## **Appendix 5.1: Practical application of linear discriminant analysis**

In order to illustrate a practical application of LDA, a data set of residential electricity consumption for the months of January and July for 1399 houses is presented in Figure 5.9. In the data set there are 362 “small” houses (average area of 115 m<sup>2</sup>) and 1037 “large” houses (average area of 250 m<sup>2</sup>). A small set of data is used from the Virtual Smart Meter data set by Neale et al. (2020a) for the purpose of this appendix.

To summarize the example in the terms presented in the paper glossary:

- *Buildings*: 1399 single-family homes.
- *Features*: January and July electricity consumption. Each building is represented by two electricity consumption values  $\{E_{jan}, E_{jul}\}$ .
- *Class*: house size, represented by categories describing the size.
- *House size categories*: ‘small’ and ‘large’.
- *Predictors*: the feature pairs  $\{E_{jan}, E_{jul}\}$  for 1399 homes.
- *Response*: the size category labels for 1399 homes, either {‘small’} or {‘large’}.

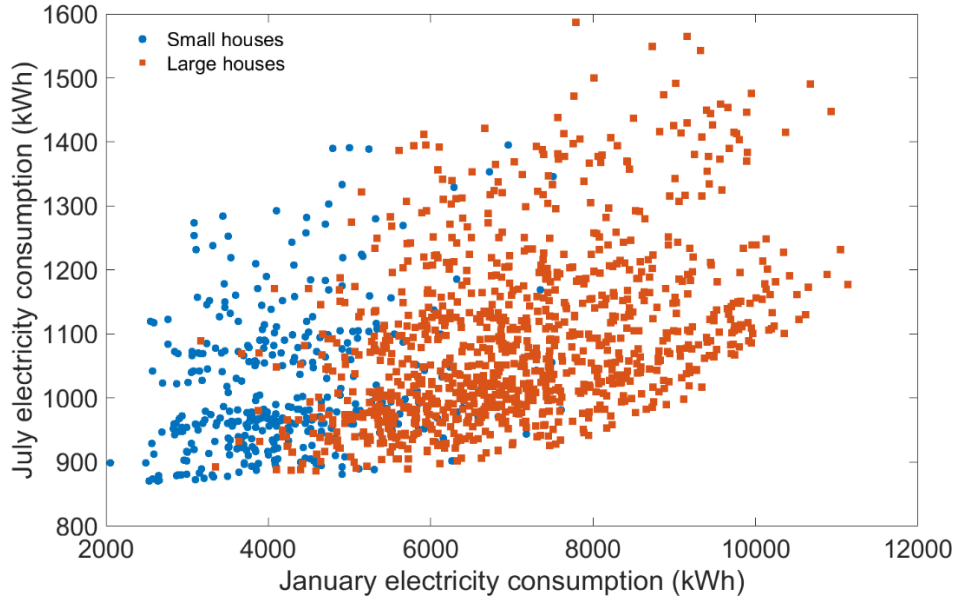


Figure 5.9. January and July electricity consumption for small and large houses. Class: house size, categories (2): small and large, features (2): January and July electricity consumption

The data in Figure 5.9 illustrates that there is some degree of overlap between the small and large house data, which is due to the variety of building parameters used to model the homes. The goal of classification using LDA would be to establish a linear decision boundary that would best separate the two data sets such that new values of January and July electricity consumption will be classified as either “small” or “large”. The probability  $p$  of a new data point belonging to one particular class category  $c$  can be expressed using Equation (5.2).

$$p_c = \frac{n_c}{n} \quad (5.2)$$

where  $n_c$  is number of samples in class category  $c$ ,  $n$  is the total number of samples. For the example given,  $p_{small} = 0.259$  and  $p_{large} = 0.741$ . Consider that the data can be divided into two subsets ( $c = 2$ ),  $\mathbf{Y}_{small}$  and  $\mathbf{Y}_{large}$ , which represent the electricity consumption data for the small houses and large houses, respectively, and can more generally be expressed as  $\mathbf{Y}_c$ . Each subset  $\mathbf{Y}_c$  also has two features ( $f = 2$ ) in this example, which can be expressed as the subsets  $\mathbf{X}_{jan}$  and  $\mathbf{X}_{jul}$ , for January and July electricity consumption, respectively. Each subset can therefore be expressed



as matrices  $[\mathbf{X}_{small,jan} \ \mathbf{X}_{small,jul}]$  and  $[\mathbf{X}_{large,jan} \ \mathbf{X}_{large,jul}]$  of size  $[n_c \times f]$ . Note that variables that are vectors or matrices are indicated with bold text.

The mean of each subset  $\mathbf{X}$  can be calculated using Equation (5.3). Since there are a number of features per class category, the resulting mean values are stored in vector form of size  $[1 \times f]$ .

$$\boldsymbol{\mu}_c = \frac{1}{n_c} \sum_{i=1}^f \mathbf{X}_{c,i} \quad (5.3)$$

where  $\boldsymbol{\mu}_c$  is a  $[1 \times 2]$  vector containing the mean values of the January and July electricity consumption values for class category  $c$ , and  $i$  is the feature count. The mean values can then be used to determine the within-class covariance, as expressed in Equation (5.4).

$$\mathbf{K}_c = \frac{1}{n_c - 1} \sum_{i=1}^f (\mathbf{X}_{c,i} - \boldsymbol{\mu}_c) (\mathbf{X}_{c,i} - \boldsymbol{\mu}_c)^T \quad (5.4)$$

where  $\mathbf{K}_c$  is the covariance matrix for class category  $c$  of size  $[f \times f]$ , or  $[2 \times 2]$  in this example. A common reduction technique used in LDA is to establish a pooled estimate of the covariance, combining the covariance matrices for the class categories. In the example given, for categories “small” and “large” houses, the pooled covariance could be expressed as in Equation (5.5).

$$\mathbf{K} = \frac{(n_{small} - 1)\mathbf{K}_{small} + (n_{large} - 1)\mathbf{K}_{large}}{n_{small} + n_{large} - 2} \quad (5.5)$$

where  $\mathbf{K}$  is the pooled covariance matrix. By using the inverse of the covariance matrix, a discriminant function  $\delta_c(x)$  can be determined, where the goal is to determine the maximum value of  $\delta_c(x)$ . For the purpose of brevity the derivation of Equation (5.6) is not presented here, but can be found in many reference texts related to machine learning techniques, such as in James et al. (2013).

$$\delta_c(x) = \mathbf{x}^T \mathbf{K}^{-1} \boldsymbol{\mu}_c - \frac{1}{2} \boldsymbol{\mu}_c^T \mathbf{K}^{-1} \boldsymbol{\mu}_c + \log(p_c) \quad (5.6)$$

where  $\mathbf{x}$  contains the feature variables, which in this case are the electricity consumption values in the months of January and July ( $E_{jan}$  and  $E_{jul}$ ), and where the mean  $\mu_c$  and the inverse matrix  $\mathbf{K}^{-1}$  are determined using Equations (5.3) and (5.5), respectively. Equation (5.6) expresses the projection of the mean and covariance of the data sets on a projection axis and establishes a decision line by maximizing the term to the right of the equal sign. The process can be completed for each class category, as illustrated in Equations (5.7) and (5.8) for the *small house* and *large house* categories, respectively.

$$\begin{aligned} \delta_{small}(x) = & [E_{jan} \quad E_{jul}] \mathbf{K}^{-1} \begin{bmatrix} \mu_{small,jan} \\ \mu_{small,jul} \end{bmatrix} \\ & - \frac{1}{2} [\mu_{small,jan} \quad \mu_{small,jul}] \mathbf{K}^{-1} \begin{bmatrix} \mu_{small,jan} \\ \mu_{small,jul} \end{bmatrix} + \log(p_{small}) \end{aligned} \quad (5.7)$$

$$\begin{aligned} \delta_{large}(x) = & [E_{jan} \quad E_{jul}] \mathbf{K}^{-1} \begin{bmatrix} \mu_{large,jan} \\ \mu_{large,jul} \end{bmatrix} \\ & - \frac{1}{2} [\mu_{large,jan} \quad \mu_{large,jul}] \mathbf{K}^{-1} \begin{bmatrix} \mu_{large,jan} \\ \mu_{large,jul} \end{bmatrix} + \log(p_{large}) \end{aligned} \quad (5.8)$$

If the discriminant functions for each category are assumed equal ( $\delta_{small}(x) = \delta_{large}(x)$ ), a linear decision boundary between those two categories can be established. By combining Equations (5.7) and (5.8) and simplifying, the resulting linear boundary separating the two class categories can be expressed as Equation (5.9).

$$h_{small:large} = \beta_{jan} E_{jan} + \beta_{jul} E_{jul} + C \quad (5.9)$$

where  $\beta_{jan}$ ,  $\beta_{jul}$  and  $C$  are constants determined from the data and where  $h_{small:large}$  is the classification rule result, which for a new coordinate of  $\{E_{jan}, E_{jul}\}$  would determine whether the data point belongs to the first or second class category. If  $h_{small:large} > 0$ , the data belongs to the “small house” category. In this specific example there are only two categories, which indicates that if the data point is not a small house, it must be a large house, but in other cases there could be multiple other categories requiring additional classification boundaries to be verified. For the example given, the coefficients are described in Equation (5.10).

$$h_{small:large} = -0.00154E_{jan} + 0.00242E_{jul} + 5.6142 \quad (5.10)$$

The linear decision boundary in Equation (5.10) can be graphed by assuming  $h_{small:large} = 0$ , which is the line of transition between the two categories. By introducing the original data set values of  $\{E_{jan}, E_{jul}\}$  into the classification boundary equation, the categories of the original data can be predicted and compared to the real category values. For the studied example, the correct and incorrect predictions are illustrated in Figure 5.10.

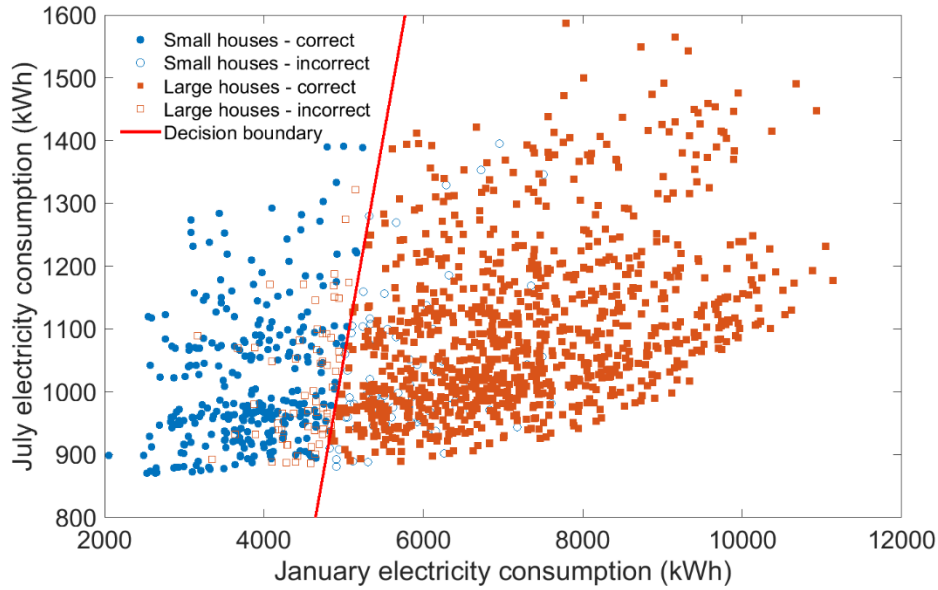


Figure 5.10. Linear classification boundary with correct and incorrect predictions

For this case, the accuracy of the classification boundary can be expressed as the correct predictions divided by the number of data points, as described in Equation (5.11).

$$CA = \frac{CP}{TP} \quad (5.11)$$

where  $CA$  is the classification accuracy,  $CP$  is the number of correct predictions and  $TP$  is the total predictions for a validation set. For this study,  $k$ -fold cross-validation is used to determine  $CP$  for all predictive models, where  $k = 5$ . Readers unfamiliar with  $k$ -fold cross-validation may refer to the end of this appendix for a description of the method.

The example presented in Figure 5.10 is simple in that it can be graphed in two dimensions, because there are only two features studied. With additional features the classification boundary becomes multidimensional, which can be expressed in a more general form as Equation (5.12).

$$h_{ci:cj} = C + \sum_{a=1}^f \beta_a E_a \quad (5.12)$$

where  $h_{ci:cj}$  is the classification rule between class categories  $ci$  and  $cj$ ,  $C$  is a constant and  $\beta_a E_a$  are the corresponding coefficients and feature values. For smart meter data, the feature values are the electricity consumption values at various moments in time that depend on the desired time aggregation, i.e. hourly, daily, monthly, etc. Equation (5.12) requires the resolution of a matrix equation that scales exponentially with the number of features and class categories, which can rapidly become quite large considering a typical year of smart meter data recorded at 15-minute intervals has 35 040 data points.

#### **Classification accuracy: $k$ -fold cross-validation**

The classification accuracy is determined by dividing the number of correct predictions (CP) by the number of total predictions (TP), as described in Equation (5.11). Establishing CP requires a data set to be divided into a test set and a validation set, commonly referred to as holdover validation. For holdover validation, the predictive model would be trained with some portion of the data set, such as 70%, and then the remaining data would be used to determine the classification accuracy. Since this prevents the use of the entire data set for training of the model,  $k$ -fold cross-validation is commonly used, which divides the data set into  $k$  segments and each segment is used to validate the predictive model developed with the remaining data, a process that is repeated  $k$  times. This allows the entire data set to be used for predictive model development and testing, but requires significantly longer computation time as the classification modeling is repeated multiple times. The classification accuracy for  $k$ -fold validation is expressed in Equation (5.13).

$$CA = \frac{\sum_{i=1}^{k_{folds}} CP_i}{TP} \quad (5.13)$$

where  $k_{folds}$  is the number of folds used for validation, which in the case of this study is 5-fold cross-validation, and  $CP_i$  is the number of correct predictions for fold  $i$ . There are 1260 correctly predicted data points out of the 1399 total data in the set. For the example illustrated in Figure 5.10,  $CA$  is equal to 0.901, i.e. 90.1% of houses are correctly classified by the illustrated linear decision boundary. An illustration of 5-fold cross-validation is presented in Figure 5.11.

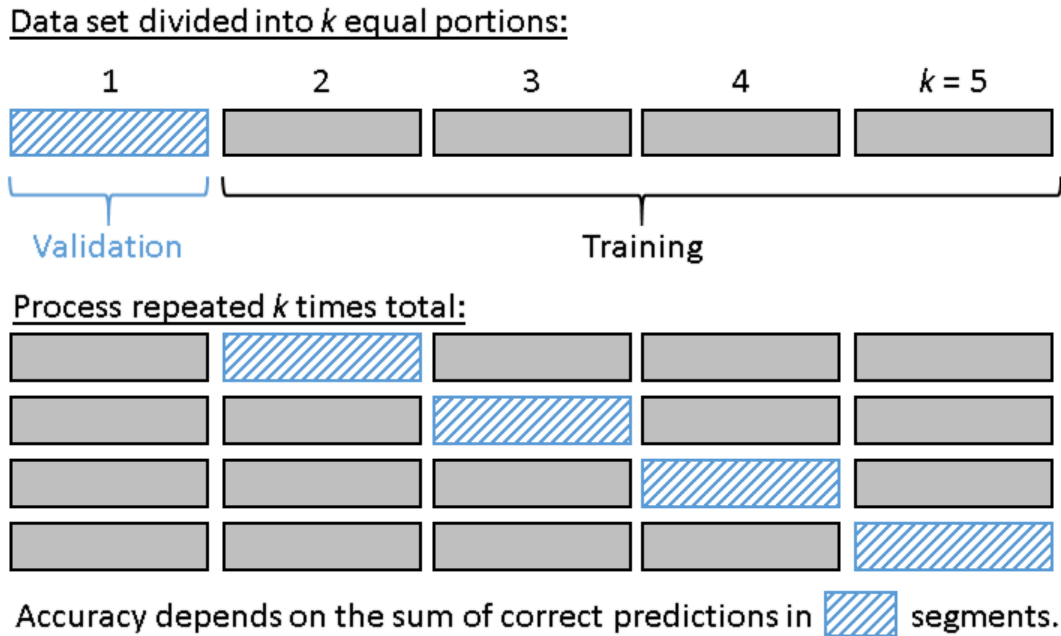


Figure 5.11. Example of 5-fold cross-validation

## **CHAPTER 6      ARTICLE 3: DEVELOPMENT OF A BOTTOM-UP WHITE-BOX BUILDING STOCK ENERGY MODEL FOR SINGLE- FAMILY DWELLINGS**

Neale, Adam, Michaël Kummert, and Michel Bernier. 2021. Submitted November 11, 2021 to the *Journal of Building Performance Simulation*.

### **6.1 Abstract**

A new bottom-up white-box building stock energy model representing 1.9 million single-family dwellings in the province of Québec, Canada, is presented in this paper. The model, called the Québec Single-Family Building Stock Energy Model (QSFBSM), is a physics simulation-based stock model that can be used to compare the base case building stock with technological variations for comparative assessments. The process of characterizing the key parameters, such as the number of dwellings in each region and the heating, cooling and hot water system distributions, is described in detail. The model accuracy is compared to known stock data for a variety of categories, including end-use, energy source and building type. The model predicts the energy use of the studied building stock with good agreement across all categories, with the total energy consumption of the model within 1.5% of the real stock energy use. The impact of the sample size of the modeled stock is evaluated, which demonstrates the importance of a sufficiently large sample to reduce the expected deviation for lesser-represented portions of the stock. A case study illustrates how the QSFBSM can be applied for a comparative assessment of different heating system distributions for the purpose of greenhouse gas emissions calculations, with an emphasis on the impact of measures on the peak electricity load of the building stock.

### **6.2 Introduction**

Residential buildings account for approximately 17% of all energy use in Canada, compared to 9% for commercial buildings (NRCan 2014). Understanding the context of the energy use for residential buildings is essential for long term planning of energy efficiency measures and technology evaluation, which has become a priority for Canada (Ugursal 2017). There are a variety

of applications for building stock modeling. For example, the provincial government of Québec has implemented a plan to reduce greenhouse gas emissions related to the heating of buildings by 50% by the year 2030 (Government of Québec 2020). There are nearly 2 million single-family dwellings in the province of Québec and non-electric energy usage represents between 22% and 42% of energy use for detached and attached homes, respectively, which is primarily for space heating and water heating (Figure 6.1). Space heating represents between 60 and 70% of the energy use of single-family homes in Québec, which has a predominantly cold climate. Reaching the Québec government's goal of 50% reduced emissions from space heating will require shifting fossil fuel energy sources to electricity and has significant implications on peak and overall electricity use.

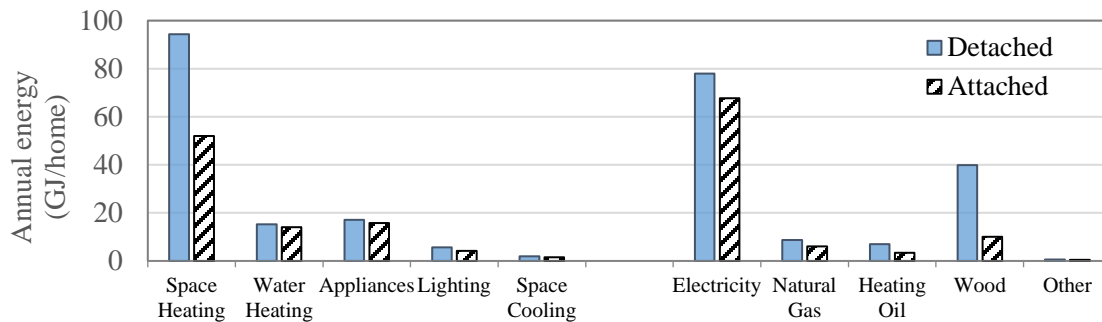


Figure 6.1 Annual energy consumption (GJ/home) by end-use and energy source for detached and attached residential dwellings in the province of Québec, Canada (NRCan 2017)

Accurately evaluating the impact of efficiency measures and technology changes on the annual and peak energy usage of homes at the municipal, regional or provincial level requires the use of some form of building stock energy model (BSEM).

### 6.2.1 BSEM approaches

Building stock energy modeling is the process of predicting the energy consumption of a large group of buildings, whether at the municipal, regional or national level. Comprehensive reviews of BSEM techniques have been performed previously by others (Swan and Ugursal 2009; Kavgić et al. 2010; Reinhart and Cerezo Davila 2016; J. Langevin et al. 2020). Methods have been typically

divided in three broad categories: top-down, bottom-up (engineering), and bottom-up (statistics) (Swan and Ugursal 2009). More recently a new classification of BSEM has emerged that distinguishes stock modeling techniques among the following “*quadrants*” (J. Langevin et al. 2020):

- Quadrant 1: Top-down black-box
- Quadrant 2: Top-down white-box
- Quadrant 3: Bottom-up black-box
- Quadrant 4: Bottom-up white-box
- Multiple quadrants (hybrid)

The distinction *top-down* establishes a building stock’s characteristics, such as energy consumption, in aggregate form and subsequently divides the stock according to various metrics, essentially starting at the so-called top level and working down to subdivided portions of the stock. *Bottom-up* models establish characteristics at the building level and subsequently aggregate the values to the desired level of accuracy, such as at the regional level. The *black-box* designation typically refers to models based on data, i.e. statistics-based models, while *white-box* refers to engineering-based cases that represent buildings and systems with physical models. A *multiple-quadrant* stock model combines several quadrants into a single approach (J. Langevin et al. 2020). Note the terms statistics-based/black-box and engineering-based/white-box will be used interchangeably in this paper.

Langevin et al. (2020) also introduce a labelling system for building stock energy models, which categorizes models based on the general purpose of the model, the quadrant, the modeling technique used to produce the results, the availability of the model, and other factors. New stock models are encouraged to be expressed in terms of those descriptors, to facilitate the comparison with other works.

Booth et al. (2012) identified the following factors as common limitations of building stock modeling:



- 1) **Accuracy:** statistics-based models can more accurately represent energy consumption in homes because they are based on real energy use, whereas engineering-based models require many assumptions.
- 2) **Data collection:** all building stock models require significant information on the stock, which is often difficult to obtain. High-resolution energy data requires significant data storage space, especially when multiple energy sources must be considered.
- 3) **Computational time:** dynamic simulations require more time to perform than statistical models, and modeling an entire stock can be prohibitive in terms of total time and resources required.
- 4) **Decision-making:** it is difficult to evaluate the propagation of uncertainty in engineering-based models, and therefore difficult to establish the impact on decisions made using the stock model.
- 5) **Flexibility:** models based on statistical data are effective as long as the base assumptions of the stock do not significantly change over time. Statistics-based models are not able to easily adapt to leaps in technology or changes in occupancy behaviour, for example.

The factors identified by Booth et al. (2012) are further supported by Langevin et al. (2020), who also describe in detail the advantages and limitations of stock modeling methods. Top-down methods either have difficulty representing technological changes due to the aggregate approach of the model, or require extensive data and an expert user to properly disaggregate the stock data into a refined model. Black-box bottom-up models require a statistical approach towards establishing building energy consumption, relying on energy use data to establish trends. As with the top-down models, relying on existing energy consumption data results in a model that is inflexible given technological advancements and relies heavily on the availability of data. A bottom-up white-box model therefore presents an interesting opportunity that can evaluate the impact of various technologies provided the computational requirements can be mitigated.

## 6.2.2 Bottom-up white-box (BU/WB) building stock modeling

While stock models can be used for various applications, energy use (and subsequently greenhouse gas emissions) is the focus of the authors. A BU/WB building stock energy model (BSEM) has a number of advantages, including but not limited to:

- 1) Energy consumption at a high frequency can be determined for each household, allowing for more accurate stock peak energy use and greenhouse gas emissions based on time-of-use.
- 2) New technologies can be implemented gradually by changing probability distributions for equipment and evaluating the impact on the stock energy use.
- 3) Occupant behaviour can be modified, such as implementing more complex time-of-use energy incentives.
- 4) Regionally-targeted incentive measures can be evaluated and compared to the rest of the building stock.

A number of BU/WB building stock energy models exist in the literature. Several residential stock modeling examples are summarized in Table 6.1 and subsequently described in more detail.

Table 6.1 Examples of recent bottom-up white-box building stock energy models

Model name	Stock	Stock size <sup>1</sup>	Building sample count (%)	Features	Market	Ref.
UMI	Urban, user-defined	~30	30 (100%)	Flexibility, customized input of an urban building stock, multiple end-use applications	Commercial, residential	(Reinhart et al. 2013)
CityBES	Urban, user-defined	10,000	10,000 (100%)	Flexibility, customized input of an urban building stock, multiple end-use applications	Commercial, residential	(Hong et al. 2016)
ResStock	National (USA)	123 million households	350,000 (0.28%)	Visualization tools, national baseline. Energy sources and end-uses by state	Residential	(Wilson et al. 2017)
AutoBEM	Urban, user-defined	130,000	130,000 (100%)	Flexibility, uses a number of imaging techniques to build 3D maps of urban settings	Commercial, residential	(New et al. 2018)
Synthetic building stock tool	National (Switzerland)	1.6 million households	10,000 (0.6%)	Auto generation of stock characteristic distributions	Residential	(Nägeli et al. 2018)
TREES	National (Japan)	53 million	16,000 (0.03%)	Detailed occupancy and house characteristics	Residential	(Taniguchi-Matsuoka et al. 2020)

<sup>1</sup> Stock size is based on the example use case provided by the authors of the tool.

The Urban Modeling Interface (UMI) allows a user to build up an urban building stock consisting of various buildings with commercial and/or residential end-uses (Reinhart et al. 2013). UMI is intended to allow a user to model any collection of urban buildings, which requires a user to input the details of the stock manually. Every building in the stock is modeled individually and therefore there is no sampling of the stock, which limits the model to smaller (urban) building stocks. It is unclear how accurate the energy prediction of UMI is, though it is based on EnergyPlus (US DOE 2013), which is a reliable energy simulation tool.

Similarly to UMI, the CityBES urban modeling software uses EnergyPlus as an engine to model a series of buildings in an urban setting (Hong et al. 2016). The user must input the details of the building stock to the model and specify the building parameters, which results in a flexible urban model requiring significant user manipulation. CityBES has a number of energy conservation measures already implemented in the tool, which allows a user to rapidly evaluate the impact of ECMs on the stock energy consumption. Much like with UMI it is unclear how accurate the energy prediction of CityBES is on a building level, though it does include an auto-calibration feature if monthly building energy consumption is available. As stated by the creators of CityBES, the computational requirements for a large urban stock, such as one million buildings in New York City, become intractable and can require a more localized urban model.

The National Renewable Energy Laboratory (NREL) has developed a national residential building stock energy model called ResStock for visualization and energy prediction of dwellings across the United States of America (Wilson et al. 2017). ResStock uses conditional probability distributions to develop residential archetypes to represent dwellings across the country. Much like UMI and CityBES, ResStock is created using the EnergyPlus simulation software. The National Baseline data viewer developed with ResStock (NREL 2021) is based on a subset of 350,000 buildings that represent approximately 123 million dwellings across the USA. The accuracy of the energy prediction across a number of housing types falls within  $\pm 20\%$  in most cases, though some significant discrepancies are identified by the creators of the tool for certain housing combinations. Overall, the ResStock tool is a very detailed example of what is possible with a BU/WB model with sufficient data on building characteristics. However, due to the size of the housing stock in

the USA, NREL must rely on a relatively small simulation (or real data) sample size (0.28%) for the aggregated energy consumption at the state and national levels.

The AutoBEM urban energy modeling tool developed by Oak Ridge National Laboratories (ORNL) again uses EnergyPlus as a basis for the prediction of urban energy consumption. The main differentiating factor for AutoBEM versus other urban tools, such as UMI and CityBES, is the use of multiple imaging techniques for the creation of 3D geometry of buildings for simulation purposes. It is difficult to evaluate the accuracy of AutoBEM as there is little detail provided by the authors on the validation of the model. An example of 130,000 mixed-use buildings in an urban setting is provided by ORNL in the form of a website (ORNL 2021).

Nägeli et al. (2018) propose a methodology and a tool based on a so-called *synthetic building stock model* using building stock characterisation and energy modeling. The energy consumption is calculated based on the monthly energy demand of each building. The tool is limited to a sample size of 10,000 buildings due to computational considerations. For the example given of the Swiss residential building stock, this results in a 0.6% sample of buildings simulated. The tool uses a detailed set of dwelling characteristics to represent each building, with a means to synthetically construct a set of dwellings and buildings based on input stock parameter distributions. While the tool has generally good agreement with building stock energy use, the tool authors acknowledge additional calibration could improve the results. In addition, given the monthly energy demand calculations it would be difficult to evaluate the impact of energy conservation measures on peak loads and other factors requiring high frequency energy consumption data.

The Total Residential End-use Energy Simulation (TREES) tool is a residential building stock energy model developed for Japan (Taniguchi-Matsuoka et al. 2020). Detailed occupancy and appliance characteristics are modeled to predict the space heating, space cooling, water heating and appliance energy consumption for randomly sampled dwellings. A simplistic thermal circuit network method is used to predict the heating and cooling loads. The sample size of the modeled stock compared to the total building stock is 0.03%, which results in significant discrepancies between reported stock energy consumption and the model prediction. Nevertheless, the structure

of the TREES model illustrates how detailed dwelling characteristics can be implemented in a BU/WB stock tool.

The illustrated BU/WB stock models provide some relevant examples of energy prediction tools for dwellings at the urban and national scales. Accuracy of the models is a recurring issue, as the sample size is limited by the scope of the stock – urban cases can be modeled entirely, while national stocks require a very small sample due to computational resource limitations. It is difficult to achieve a high degree of accuracy when less than 0.1% of homes are simulated. Few details, if any, are provided on the impact of the sample size of BU/WB tools on overall accuracy. In addition, stock energy models appear to be mainly focused on energy consumption when peak loads are often an important consideration for stakeholders. Most tools are used for comparative studies, such as evaluating the impact of an energy conservation measure, in which case the absolute accuracy of the tool seems to be secondary to the tool authors.

### **6.3 Objectives**

This paper describes a new bottom-up white-box building stock energy model (BSEM), called the Québec Single-Family Building Stock Energy Model (QSFBSSEM). The objective of the QSFBSSEM is to provide a validated stock model that can evaluate different technology and building stock scenarios and study the impact on energy usage and peak loads for a variety of energy sources and end-uses. While the general methodology presented in this paper is applicable to other building stocks, the authors have applied it to the single-family dwelling market in the province of Québec, a building stock representing 1.9 million homes. More specifically, this paper aims to:

- 1) Describe the characterisation process of the studied building stock, including a description of typical dwellings, population distribution, climate zones and common building systems.
- 2) Present a general methodology to develop a building stock model from a series of building energy simulations of individual houses.
- 3) Describe the characterization and implementation of the stock model in detail.
- 4) Establish the impact that sample size has on building stock energy prediction.

- 5) Validate the aggregate results of the proposed model with building stock energy consumption data.
- 6) Present an example application of the stock model.

This paper builds on a previous work by Neale et al. (2020a) that presented a methodology to develop a virtual smart meter data set. Some of the building stock characterisation was presented previously, though the work by Neale et al. was for electricity use profiles only and did not consider other energy sources. In order to reduce repetition some details are summarised here and readers can refer to the paper by Neale et al. for further details. In many cases, data and probability distributions for characteristics have been updated, or additional details were added, and therefore are presented in this paper for clarity.

## 6.4 Building stock characterisation

The segmentation (sometimes called classification) and characterisation processes are often used for building archetype development (Sokol et al. 2016), which is a bottom-up white-box technique commonly used for building stock modeling (Swan and Ugursal 2009; J. Langevin et al. 2020). Segmentation is the process of determining the parameters that differentiate different types of buildings, such as climate zones, house types, etc. Characterisation is used to identify the range of values of each parameter given the building stock composition, such as detached and semi-detached houses for the *house type* parameter.

The characterisation process is applied to single-family dwellings in the province of Québec to generate accurate houses that fit the range of parameters found in the building stock. Characteristics of the housing stock, such as the types of dwellings, number of dwellings across the province, heating and cooling systems are described in detail. The data collected serves to establish probability distributions used to generate combinations of parameters that exist within the building stock.

### 6.4.1 Dwelling types

The province of Québec, Canada, is characterized by a residential building stock consisting of a mix of single-family dwellings (SFD) and multi-residential dwellings (MRD). From a modeling

point of view there are distinct differences between SFD and MRD, such as the boundary conditions above and below a detached home are significantly different from an apartment in a multi-residential building. This study aims to develop a stock model for SFD only, and therefore targets the following types of homes:

- Single-detached homes (Det)
- Semi-detached homes (Semi)
- Row houses (Row)
- Other single-attached (OSA)

The four building types in the list above, depicted in Figure 6.2, are commonly used by national statistics and energy data publications in Canada to present dwelling distributions and energy consumption data. Occasionally reference data is presented for “Detached” and “Attached” homes, in which case Semi, Row and OSA homes are combined for the latter category as they all share external boundaries with adjacent buildings.

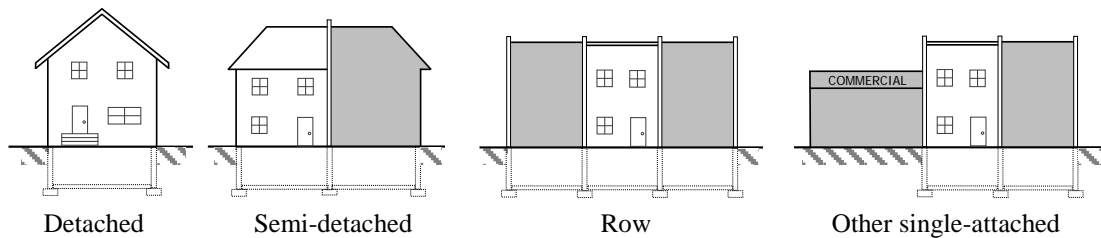


Figure 6.2 Single-family dwelling (SFD) types in the province of Québec, Canada

Houses in the studied building stock are most commonly one or two-storeys plus a heated basement. The number of storeys is dependent on the type of dwelling, with Semi, Row and OSA more commonly having two-storeys (NRCan 2018). The heated surface area of the dwelling also depends on the building type and number of storeys.

#### 6.4.2 Number of dwellings by region

There is a variety of sources for data related to the number of residential dwellings in the Province of Québec, Canada. The Canadian Census of Population Program (CCPP) from Statistics Canada is a reliable source for *Type of Dwelling* data since 2016, where census responders indicate relevant

characteristics of the home that they live in (Statistics Canada 2016). The CCPP divides Canada into a number of Census Metropolitan Areas (CMA), which are population hubs of more than 100,000 people of one or more neighbouring municipalities where at least 50,000 individuals live in the urban core, and Census Agglomerations (CA), which have a core population of at least 10,000 (Statistics Canada 2016). The province of Québec has six CMA regions and twenty-four CA regions, which are illustrated on the population density map in Figure 6.3 (Statistics Canada 2019).

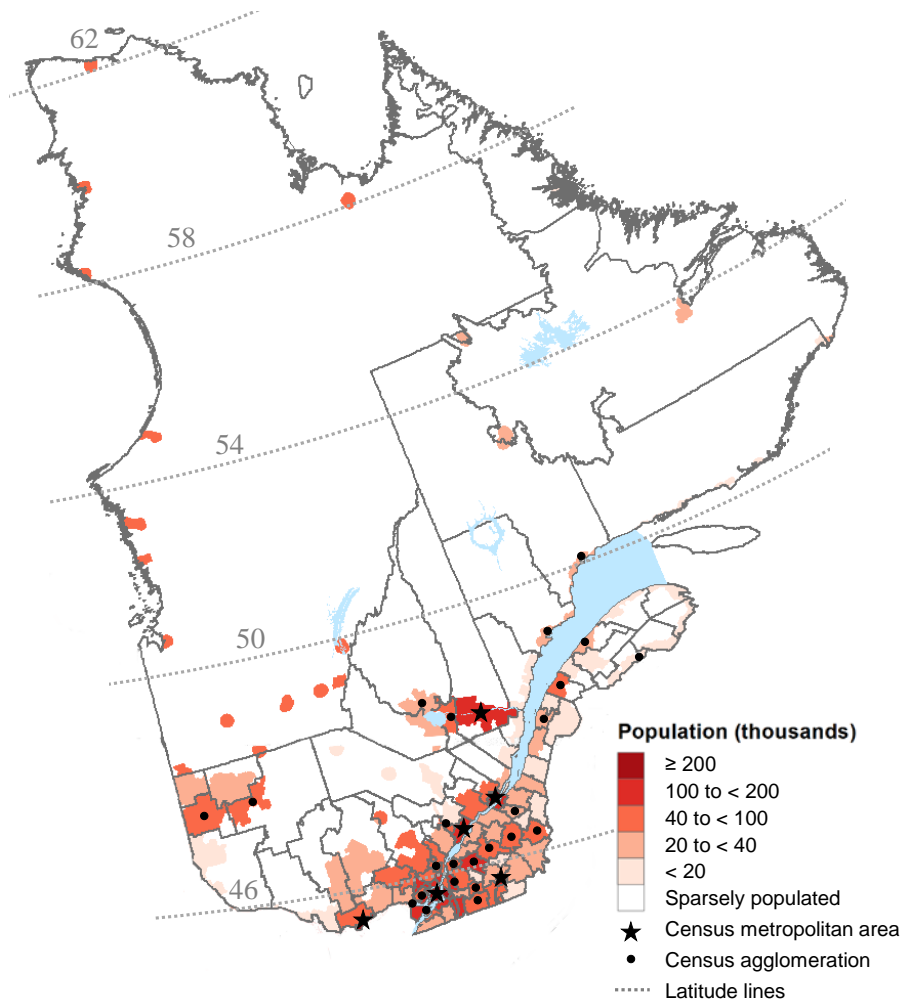


Figure 6.3 Population distribution of the province of Québec, Canada, with CMA and CA regions superimposed. Adapted from Statistics Canada (2019). Approximate latitude lines are indicated for reference



The six CMA regions represent significant concentrations of the population and, consequently, of dwellings. Energy use data for the province of Québec is often expressed in terms of these CMA regions, with a seventh region entitled “Québec non-CMA”, which represents the remainder of the province (NRCan 2015). The distribution of single-family dwellings (SFD) across these seven regions is presented in Table 6.2 (Statistics Canada 2016).

Table 6.2 Distribution of occupied dwellings for Québec CMA areas. DF: dwelling fraction, SFD: single-family dwelling, Det: single-detached house, Row: row house, Semi: semi-detached house, OSA: other single-attached house (Statistics Canada 2016)

Regions	Number of dwellings by region					Dwelling fraction (DF) by region				
	SFD	Det	Row	Semi	OSA	DF <sub>SFD</sub>	DF <sub>Det</sub>	DF <sub>Row</sub>	DF <sub>Semi</sub>	DF <sub>OSA</sub>
R1 Québec Non-CMA <sup>1</sup>	780,600	670,777	22,342	78,255	9,225	0.4108	0.8593	0.0286	0.1003	0.0118
R2 Saguenay	44,195	36,900	5,455	1,580	260	0.0233	0.8349	0.0358	0.1234	0.0059
R3 Québec City	180,380	148,965	21,135	8,935	1,345	0.0949	0.8258	0.0495	0.1172	0.0075
R4 Sherbrooke	49,780	42,630	4,320	2,455	375	0.0262	0.8564	0.0493	0.0868	0.0075
R5 Trois-Rivières	40,710	33,720	4,785	1,845	360	0.0214	0.8283	0.0453	0.1175	0.0088
R6 Montréal	713,710	564,230	86,460	56,770	6,250	0.3756	0.7906	0.0795	0.1211	0.0088
R7 Gatineau	90,845	64,710	17,890	7,955	290	0.0478	0.7123	0.0876	0.1969	0.0032
Total SFD	1,900,220	1,602,675	188,245	93,355	15,945					

<sup>1</sup> The dwelling fraction values for Québec non-CMA are extrapolated from the number of houses for the 24 census agglomerations regions, i.e. 230,000 dwellings are extrapolated to represent 780,600.

The seven identified regions represent 100% of the single-family dwellings in the province of Québec, Canada, or approximately 1.9 million houses (2017 data). The value of  $DF_{SFD}$  represents the fraction of all single-family dwellings in each region. The following columns,  $DF_{Det}$  to  $DF_{OSA}$ , represent the corresponding fraction of dwellings of that type for that region. As an example, 9.49% of single-family dwellings are located in R3 Québec City, and of those 82.6% are detached houses.

The 24 CA regions represent approximately 230,000 dwellings, which is 31% of the Québec non-CMA amount of 780,600 (Statistics Canada 2016). The remainder of the homes are located in less densely populated areas in relative proximity to the CA and CMA regions or in small settlements

in the far north of the province. For the purpose of this study, the twenty-four CA regions identified in Figure 6.3 are considered to provide an accurate representation of the Québec non-CMA portion of the building stock, as they are the highest concentration of buildings outside of the metropolitan areas. The distribution of homes in the 24 CA regions is therefore extrapolated to be equal to the total number of homes in the remainder of the province. The description of the 24 sub-regions of R1 is provided in Appendix 6.1. The results of the 24 sub-regions are aggregated and referred to as *Québec non-CMA* for the remainder of this study.

### 6.4.3 Weather data

Building energy models often use typical weather files to standardize the heating and cooling loads and to remove the annual fluctuation in conditions found in real weather. The Canadian Weather year for Energy Calculation (CWEC) files serve that purpose for Canadian regions (Government of Canada 2021). CWEC files contain 12 distinct typical meteorological months that are selected from a database of 30 years of weather data. Each combination of months is distinct for each region, which results in months selected from different years of data for each weather file. When modeling a building stock with multiple climate zones, non-coincidental peaks pose an issue as the peak energy use does not occur at the same moment for each region. As an example, the outdoor dry bulb temperature from CWEC files for the six CMA regions are illustrated in Figure 6.4 (left) for the month of December. Each region is identified with the corresponding year that the month of December is taken from, i.e. R2-2009 indicates the calendar year 2009 for region R2.

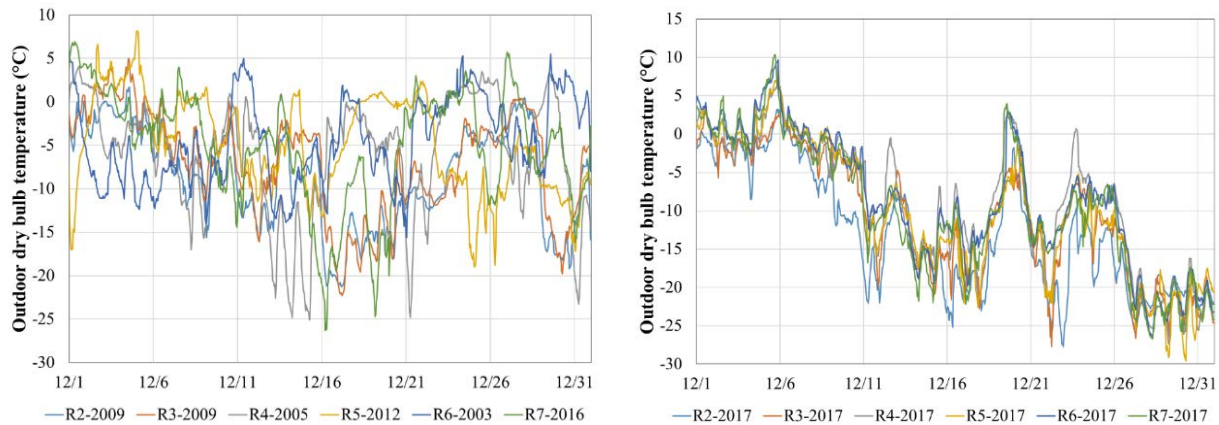


Figure 6.4 Outdoor dry bulb temperature for 6 CMA locations across the province of Québec for CWEC weather data (left) and 2017 CWEEDs data (right)

The temperature curves in Figure 6.4 (left) appear to be independent, except for R2 and R3, which both use the 2009 data for the month of December. For the purpose of comparison, the Canadian Weather Energy and Engineering Datasets (CWEEDs) data for 2017 is depicted in Figure 6.4 (right) for the same six CMA regions of the province (Government of Canada 2021).

While there are regional differences, the general temperature trend is the same across the six illustrated regions for the CWEEDs data. For the purposes of modelling a building stock's energy consumption and peak load, a set of coincident weather data for the relevant regions is necessary for accurate prediction of the peak heating load. In addition, there is energy consumption data for validation purposes for specific years. After comparing multiple years of weather data, the calendar year 2017 is selected for the results presented in this paper, though the model can function with any year of weather data. For comparison purposes, the heating degree-days (HDD@18°C) for the 2017 CWEEDs data vary between approximately 4000 and 5400 degree-days across the studied building stock.

#### 6.4.3.1 Weather for R1: Québec Non-CMA

Aside from the six census metropolitan areas, which represent a significant portion of the building stock, a seventh region is required to represent the remainder of the homes in the province. The 24 census agglomeration regions are selected to divide the remaining buildings into populated regions

spread across Québec, as illustrated in Figure 6.3. The closest weather station to each CA region was identified geographically. A house in *Québec non-CMA* is therefore assigned to one of the 24 weather stations according to the probability distribution described in Table 6.10 in Appendix 6.1.

#### **6.4.3.2 CWEEDs missing cloud cover data**

In the process of analysing the CWEEDs weather data for the 30 regions of this study it was discovered that Environment Canada no longer recorded Total Sky Cover or Opaque Sky Cover for most weather stations across the province of Québec from 2013 onwards. In some cases, such as at certain airports, Total Sky Cover was recorded but only at 3 hour intervals. The TRNSYS building energy simulation software used in this study requires the Opaque Sky Cover in order to determine the sky temperature for longwave radiation calculation (Klein et al. 2017). In order to correct this issue a methodology is applied to fill the Total and Opaque Sky Cover values in the weather files used for the building stock model, which is presented in Appendix 6.2.

#### **6.4.4 Building performance characteristics**

The construction year of a home is not a strong indicator of thermal performance characteristics for the studied building stock. Building performance is therefore better characterized directly by wall, roof, foundation thermal resistance, window type and leakage area, as opposed to by construction vintage. Probability distributions for house characteristics related to thermal performance are determined from the Energuide Housing Database containing over 27,000 homes in the studied building stock (NRCan 2018).

#### **6.4.5 System characterization**

The single-family residential building stock studied in this paper relies on a variety of space heating, space cooling and water heating technologies. The characterization process for these systems is described in the following sections of the paper, which will be used for modeling purposes in the building energy simulation program.

### 6.4.5.1 Space heating

The data related to the prevalence of different heating technologies for attached and detached dwellings for the studied building stock is presented in Table 6.3, which originates from the Canadian Comprehensive Energy Use Database (CEUD) (NRCan 2017). Systems representing less than 1% of the building stock are excluded, whether for attached or detached dwellings. Heating systems are labelled from H1 to H8.

Table 6.3 Heating system distribution for the Province of Québec for single detached and single attached homes (NRCan 2017)

Primary heating system description		Detached		Attached	
		#units (thousands)	Probability	#units (thousands)	Probability
H1	Heating Oil – Medium Efficiency	134.5	0.078	28.0	0.085
H2	Natural Gas – High Efficiency	52.1	0.030	10.8	0.033
H3	Electric	741.3	0.432	227.4	0.689
H4	Heat Pump	212.3	0.124	16.6	0.050
H5	Wood/Electric	370.3	0.216	10.5	0.032
H6	Wood/Heating Oil	94.9	0.055	4.1	0.012
H7	Heating Oil/Electric	111.4	0.065	16.2	0.049
H8	Wood	0.0	0.000	16.5	0.050

The heating system data from (NRCan 2017) provides an average distribution across all regions for the studied building stock. There are significant differences between the detached and attached heating system distributions, most notably in the increased prevalence of electric baseboard heating (H3) in attached houses and the lack of wood-heated (H8) detached houses. In addition, the Survey of Household Energy Use of 2015 includes data on primary heating system energy type by region, illustrated in Table 6.4 (NRCan 2015).

Table 6.4 Fraction of homes by region based on primary heating energy (NRCan 2015)

	Electricity	Natural gas	Wood	Unknown <sup>1</sup>
R1 Québec Non-CMA	0.763	-	0.178	0.060
R2 Saguenay	0.920	-	-	0.080
R3 Québec City	0.870	-	-	0.130
R4 Sherbrooke	0.830	-	-	0.170
R5 Trois-Rivières	0.897	-	-	0.103
R6 Montréal	0.882	-	-	0.118
R7 Gatineau	0.289	0.623	-	0.088

<sup>1</sup> Data missing or unaccounted for

While there are significant gaps in the SHEU data set in Table 6.4, there are several important details that can be used to complement the data presented in Table 6.3. For example, there is a prevalence of natural gas heating in *R7 Gatineau* that is uncharacteristic of the remainder of the province, and wood heating is more common in the outlying regions of *R1 Québec Non-CMA*. Electricity-based systems are considered to be distributed according to the *electricity* fraction in Table 6.4. The *unknown* column represents missing data and is presumed to belong to heating systems not represented in the included data. As an example, for R2 Saguenay the 0.08 fraction of missing data is distributed among systems without electric primary heating, i.e. H1, H2, H6 and H8. The combination of the probability distributions in Table 6.3 and Table 6.4 result in regional probability tables for the heating systems of detached and attached buildings in Table 6.5.

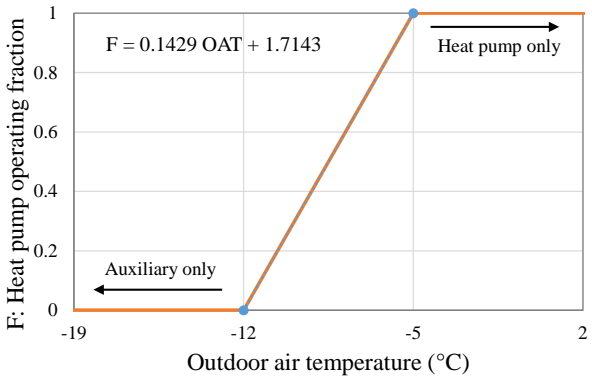
In order to illustrate how the values in Table 6.5 are determined, consider the case for R1 (*Québec Non-CMA*) for single-detached homes. The fraction of electric-heated homes for R1 is 0.763 from Table 6.4. Heating systems using electricity (H3, H4, H5 and H7) are distributed based on the weighted distribution in Table 6.3. Wood systems represent a fraction of 0.178, though only hybrid wood systems are found in detached homes for the studied building stock. Therefore the wood/heating oil systems are assumed to represent 0.178 of systems in R1, since wood/electric heating is considered primarily an electric system and there are no wood-only systems (H8) in R1. The remaining heating systems (H1 and H2) are therefore distributed according to the *unknown* fraction 0.060 from Table 6.4 according to their respective probabilities.

Table 6.5 Heating system probability by region. Highlighting by data source from Table 6.4: electric, natural gas, wood or unknown.

Region	Heating system fraction by region							
	H1 Heating oil	H2 Natural gas	H3 Electric	H4 Heat pump	H5 Wood/ Electric	H6 Wood/ Heating oil	H7 Heating oil/ electric	H8 Wood
Single detached								
R1	0.043	0.017	0.394	0.113	0.197	0.178	0.059	0.000
R2	0.038	0.015	0.475	0.136	0.237	0.027	0.071	0.000
R3	0.062	0.024	0.450	0.129	0.225	0.044	0.068	0.000
R4	0.081	0.032	0.429	0.123	0.214	0.057	0.064	0.000
R5	0.049	0.019	0.463	0.133	0.231	0.035	0.070	0.000
R6	0.056	0.022	0.456	0.130	0.228	0.040	0.068	0.000
R7	0.052	0.623	0.149	0.043	0.074	0.036	0.022	0.000
Single attached								
R1	0.043	0.017	0.641	0.047	0.030	0.035	0.046	0.142
R2	0.038	0.014	0.773	0.056	0.036	0.006	0.055	0.022
R3	0.061	0.024	0.731	0.053	0.034	0.009	0.052	0.036
R4	0.080	0.031	0.697	0.051	0.032	0.012	0.050	0.047
R5	0.049	0.019	0.754	0.055	0.035	0.007	0.054	0.029
R6	0.056	0.021	0.741	0.054	0.034	0.008	0.053	0.033
R7	0.051	0.623	0.242	0.018	0.011	0.007	0.017	0.030

The eight identified heating systems operate according to the descriptions provided in Table 6.6. Single energy source systems are assumed to cover the entire heating load of a dwelling. Hybrid systems are attributed specific fractional loads based on the prevalence of those types of systems in the studied province. For example, favourable electricity rates are provided to homeowners with hybrid electric and heating oil systems when the outdoor temperature is below -12°C (Hydro-Québec 2021), and therefore these types of systems are assumed to transition between systems at low temperatures.

Table 6.6 Detailed heating system descriptions. O: heating oil, NG: natural gas, E: electric, W: wood. OAT: Outdoor air temperature

Heating system		Description	System details	Secondary electricity use	System efficiency
H1	Heating Oil – Medium Efficiency	Heating oil boiler or furnace	Heating load 100% covered by heating oil-fired system.	2.39% <sup>1</sup>	O: 0.78 <sup>2</sup>
H2	Natural Gas – High Efficiency	Natural gas boiler or furnace	Heating load 100% covered by natural gas-fired system.	2.39% <sup>1</sup>	NG: 0.90 <sup>2</sup>
H3	Electric	Electric baseboard or electric furnace	Heating load 100% covered by electric heating element system, i.e. baseboard heating or electric furnace.	N/A	E: 1.00 <sup>2</sup>
H4	Heat Pump	Air-source heat pump with auxiliary electric element	<p>OAT above -5 °C: Heating load 100% covered by heat pump.</p> <p>OAT below -12 °C: Heating load 100% covered by electric heating elements.</p> <p>OAT between -5 °C and -12 °C: linear fraction transitioning between the two systems as a function of temperature.</p>	N/A	<p>E: 1.00<sup>2</sup></p> <p>Heat pump<sup>3</sup>: COP = 0.0585 OAT + 3.115</p>
 <p style="text-align: center;">F: Heat pump operating fraction</p> <p style="text-align: center;">Outdoor air temperature (°C)</p> <p style="text-align: center;"><math>F = 0.1429 \text{ OAT} + 1.7143</math></p> <p style="text-align: center;">Auxiliary only      Heat pump only</p>					
H5	Wood/Electric	Electric baseboard or electric furnace with wood stove or fireplace	<p>OAT below 0 °C: wood-fired system (i.e. fireplace or stove) contributes up to 40,800 kJ/h for detached houses and 23,800 kJ/h for attached houses to the heating load, with the remainder covered by electric heating elements.<sup>1</sup></p> <p>OAT above 0 °C: heating load 100% covered by electric heating elements.</p>	N/A	<p>W: 0.50<sup>2</sup></p> <p>E: 1.00<sup>2</sup></p>



Heating system	Description	System details	Secondary electricity use	System efficiency
H6	Wood/Heating Oil	Heating oil boiler or furnace with wood stove or fireplace OAT below 0 °C: wood-fired system (i.e. fireplace or stove) contributes up to 40,800 kJ/h for detached houses and 23,800 kJ/h for attached houses to the heating load, with the remainder covered by heating oil system. OAT above 0 °C: heating load 100% covered by heating oil system.	2.39% <sup>1</sup>	W: 0.50 <sup>2</sup> O: 0.78 <sup>2</sup>
H7	Heating Oil/Electric	Heating oil boiler or furnace with electric baseboard or furnace Considered a hybrid system following hybrid electricity rate operation <sup>4</sup> , transitioning at low outdoor temperatures to non-electric systems. OAT below -12 °C: heating load 100% covered by heating oil system. OAT above -12 °C: heating load 100% covered by electric heating elements.	2.39% <sup>1</sup>	O: 0.78 <sup>2</sup> E: 1.00 <sup>2</sup>
H8	Wood	Wood stove, fireplace or furnace Heating load 100% covered by wood-fired system.	N/A	W: 0.50 <sup>2</sup>

<sup>1</sup> Based on 7 kWh/MMBtu average (NYSERDA 2013)

<sup>2</sup> NRCan (2017b)

<sup>3</sup> Johnson (2013)

<sup>4</sup> Hydro-Québec (2021)

### 6.4.5.2 Space cooling

The presence of an air conditioning (AC) system varies by region according to the Survey of Household Energy Use of 2015 (NRCan 2015). The SHEU data set has data indicating the presence of a central air conditioner, window air conditioner or no air conditioner. However, since the data for homes with central versus window AC is unreliable or missing, the probability is expressed simply as whether AC is present or not for each region.

Table 6.7 Air conditioning prevalence in the studied building stock (NRCan 2015)

Region	AC	No AC
R1 Québec Non-CMA	0.376	0.624
R2 Saguenay, Québec	0.539	0.461
R3 Québec, Québec	0.366	0.634
R4 Sherbrooke, Québec	0.596	0.404
R5 Trois-Rivières, Québec	0.630	0.370
R6 Montréal, Québec	0.680	0.320
R7 Ottawa-Gatineau, Ontario/Québec	0.879	0.121

In addition, if a home has a heat pump heating system (H4), the house is presumed to have air conditioning as most residential heat pumps are reversible. Finally, the average Energy Efficiency Rating (EER) of cooling systems in the province of Québec is EER=14 (NRCan 2017), which is applied as a constant coefficient of performance (COP=4.1).

It should be noted that the data from SHEU for air conditioning is for all residential dwellings, including multi-residential apartment buildings. Other data sources related to air conditioning exist, but the SHEU data is retained because it differentiates the probability of a cooling system by region, which has a significant impact on the distribution of air conditioners in the province.

### 6.4.5.3 Water heating

Water heating for domestic purposes is predominantly electric for the studied building stock. The probability of a home's domestic hot water (DHW) energy source is presented in Table 6.8 (NRCan 2017). Water heating systems using the *Other* category are negligible for the studied building stock and are not represented in the model.

Table 6.8 Domestic hot water system distribution by energy source (NRCan 2017)

<b>Water heater type</b>	<b>P<sub>DHW</sub></b>
Electric	0.930
Natural gas	0.042
Oil	0.021
Other	0.007

The domestic hot water fuel type is dependent on the primary heating energy source. Homes with primary electric heating systems are 99.2% likely to have electric water heating (NRCan 2018). Dwellings with non-electric primary heating often have non-electric DHW systems, though this is not universally true. Homes with natural gas-based heating systems have a 66.1% probability of having natural gas water heating, and houses with heating oil-based heating systems have 22.5% probability of having heating oil domestic hot water heating (NRCan 2018).

### 6.4.6 Occupancy

The number of occupants for the studied building stock is based on the type of building, which in turn depends on the region where the house is located. Due to the limitations of the stochastic occupancy model used to represent the internal loads of each house, the maximum number of occupants is set to 5 (McKenna and Thomson 2016). Additional details on the modeling of internal loads are presented in Section 6.5.1 of this paper.

## 6.5 Model description

The proposed building stock energy model uses an approach that generates individual dwellings according to the characterisation process described in Section 6.4 of the paper. Parameters are generated according to interdependent probability distributions combined using Bayes' Theorem. In brief, where applicable, conditional probability distributions are generated in order to produce configurations of building parameters common to the studied building stock. The result is a set of buildings that match the studied building stock.

The modeling process is first described briefly, with additional details located in an appendix to the paper. The overall accuracy of the stock model is then presented with respect to known stock

energy consumption. Next, some observations are provided about the stability of the stock energy results with respect to the sample size used to represent the building stock.

### **6.5.1 Building energy simulation**

The TRNSYS simulation program is used to model the energy use of each dwelling in the studied building stock (Klein et al. 2017). Each house is modeled individually by generating unique, randomly generated sets of building parameters according to the probability distributions established for every modeled component of the building. A user can execute a simulation for a variety of stock sample sizes and expect that the correct distribution of house parameters will be applied by the model, with some statistical variation. The overall methodology behind the operation of the model is provided in Appendix 6.3 for interested readers.

### **6.5.2 Stock energy consumption comparison**

The energy consumption by end-use and energy source is available for detached and attached buildings in the studied building stock (NRCan 2017). The end-use values include space heating, water heating, appliances, lighting and space cooling, while energy sources include electricity, natural gas, heating oil and wood. The single-family dwelling building stock combines detached and attached dwellings to include approximately 1.9 million homes in total, which are represented by a sample of 200,000 simulated dwellings. Considering the three dwelling categories and ten energy consumption values, this provides a total of 30 points of comparison for the building stock energy model, which are illustrated in Figure 6.5. Due to the good agreement of the QSFBSSEM model with the stock data, the authors felt that no calibration was necessary.

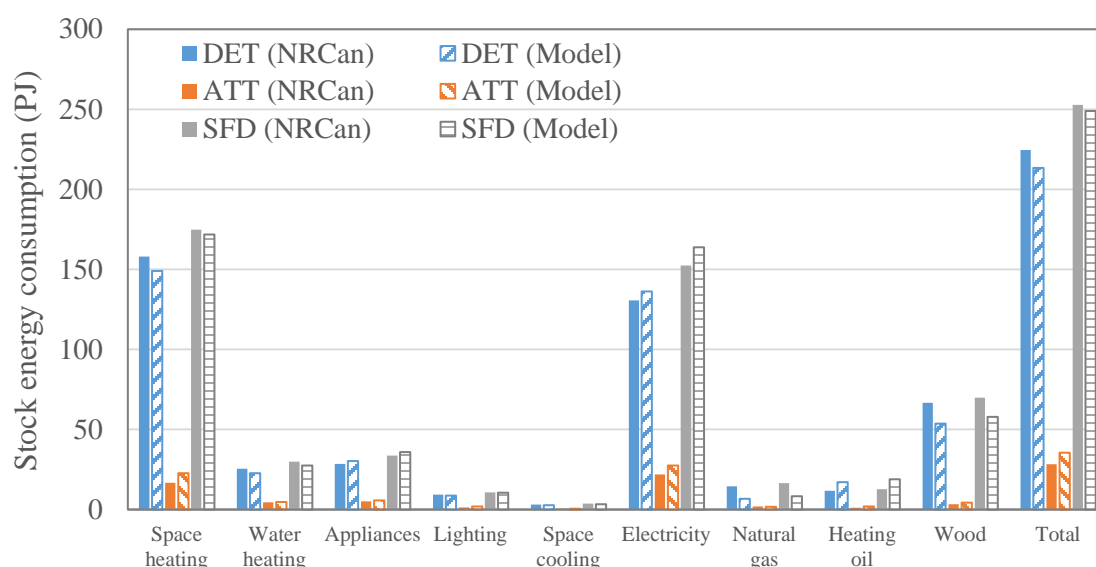


Figure 6.5 Model versus stock energy consumption for detached (DET), attached (ATT) and all single-family dwellings (SFD)

The total stock energy consumption values are illustrated in petajoules in Figure 6.5. There is good agreement for the end-use categories (space heating, water heating, appliances, lighting and space cooling). There are some differences in the natural gas, heating oil and wood categories, which predominantly contribute to space heating and water heating. Overall, the model provides total stock energy consumption within 1.58% of the stock data (NRCan value of 252.9 PJ compared to the model value of 248.9 PJ). In order to visualize these differences further, the modeled per-house energy consumption for all single-family dwellings is illustrated as a box and whisker plot in Figure 6.6.

The box plot in Figure 6.6 illustrates the interquartile range for the energy consumption by category for the 200,000 homes of the building stock energy model sample. Outliers are indicated as red data points. Most houses with natural gas and heating oil are considered outliers due to the fact that only a small percentage of homes are heated with those systems, and therefore most houses in the studied stock have zero natural gas and heating oil energy consumption. As a further example, the *Space heating* box plot illustrates the mean space heating energy consumption of the SFD stock as

a black X symbol, the mean space heating energy according to known stock data (NRCan 2017) as a semi-transparent blue circle.

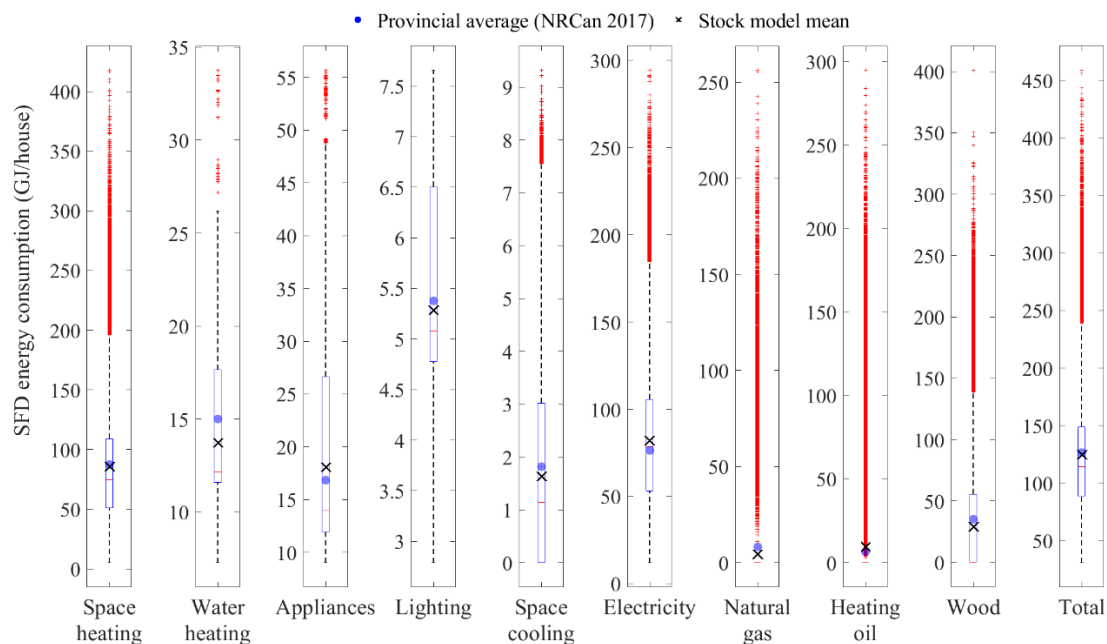


Figure 6.6 Box and whisker plot of the modeled single-family dwelling (SFD) energy consumption by end-use and energy source

Overall, the proposed building stock energy model has good agreement across the 30 studied end-use and source energy consumption categories. Further improvements could be made in the future to adjust the heating system distributions and/or refine the validation energy consumption values as additional information becomes available. Given that no model calibration is performed to improve the fit of the results to the NRCan stock data, the authors are satisfied that the proposed building stock energy model is a close representation of single-family dwellings in the province of Québec, Canada.

### 6.5.3 Building stock sample size

Much like previous works, this study takes the approach of selecting a smaller subset of homes and scaling up the energy consumption of that subset. In such cases, an important aspect of building stock energy modeling is determining the size of the sample set of homes that accurately represents

the overall stock. The impact of the sample size is expressed in terms of the normalized root mean square deviation (NRMSD) with respect to the mean of a sample. More precisely, the following procedure is followed:

1. The stock sample size  $b$  is selected, e.g. 5000 buildings, which will represent the building stock of size  $s$ , i.e. 1.9 million buildings for single-family dwellings.
2. The stock fraction  $SF$  is determined, where  $SF = b/s$ .
3. A building stock sample of size  $b$  is generated with the proposed model and the total energy consumption is determined for a variety of end-use and energy source categories. This process is repeated  $n$  times, which is up to 4000 repetitions depending on the sample size.
4. The size of the building stock sample is increased and steps (1) to (3) are repeated.
5. The  $NRMSD$  for each tested stock sample size is determined, as described in Equation (6.1).

$$NRMSD = \frac{\sqrt{\frac{\sum_{i=1}^n (E_i - \bar{E})^2}{n}}}{\bar{E}} \quad (6.1)$$

where:

- $NRMSD$  is the normalized root mean square deviation for a given sample size
- $E_i$  is the calculated stock energy for sample of buildings  $i$  (PJ)
- $\bar{E}$  is the mean stock energy across all samples of size  $b$  (PJ)
- $n$  is the number of total sets of stock samples

The  $NRMSD$  values are illustrated in Figure 6.7 as a function of total energy consumption and building type (attached, detached and single-family dwellings) for a variety of end-use and energy source categories. The  $NRMSD$  metric demonstrates the expected residual on the total energy consumption for a given stock fraction.

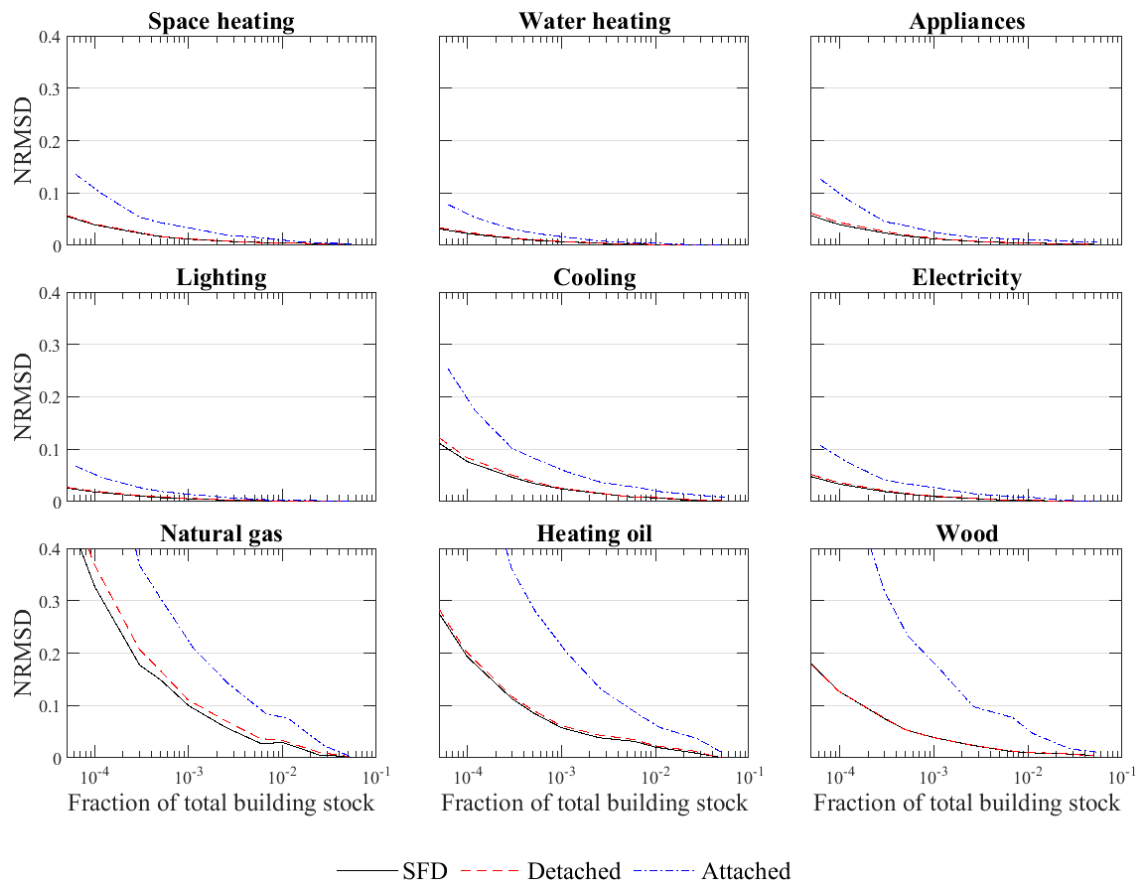


Figure 6.7 NRMSD for energy consumption by end-use and energy source by fraction of the total building stock modeled

The NRMSD of the energy consumption varies considerably depending on the stock fraction, building type and energy category. Categories strongly tied to electricity consumption have lower residuals than nonelectric categories, which is due to the prevalence of electricity in the studied building stock. Lesser-represented categories, such as natural gas, heating oil and wood, are more disposed to variations in the total energy use for smaller stock samples. Attached houses experience larger NRMSD values than detached houses for the same reason. As the fraction of the building stock approaches 10%, the NRMSD is well below 0.01 for all energy categories.

For a more specific comparison, the literature shows that stock model samples of 0.03% (Taniguchi-Matsuoka et al. 2020) and 0.6% (Nägeli et al. 2018) have been applied in previous works. For the building stock studied here, the NRMSD is illustrated for the same stock sample



sizes used by Taniguchi-Matsuoka et al. and Nägeli et al. in Figure 6.8. The same 10 energy use categories presently previously are depicted.

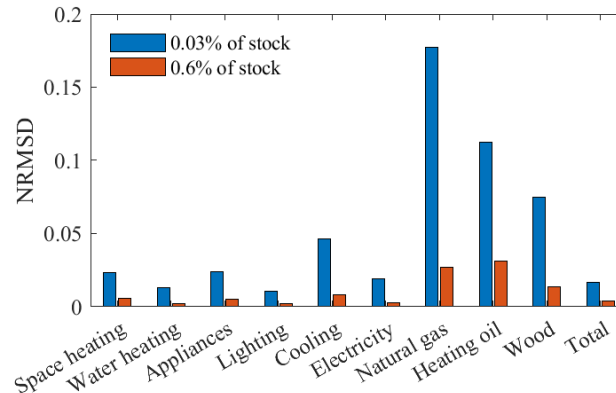


Figure 6.8 NRMSD for SFD energy consumption categories for cases representing 0.03% and 0.6% of the modeled building stock

The NRMSD is generally under 0.03 for the 0.6% stock sample, but can reach 0.18 for a sample of 0.03% of stock. In simpler terms, a deviation of 18% on the natural gas energy prediction is expected for a stock sample of 0.03% for the studied building stock. The *Total* energy use NRMSD is low in both cases, but if specific categories are required for a particular analysis then the resulting deviation can be quite large. For example, if the stock model is used to evaluate cooling energy savings due to a particular incentive measure, the NRMSD can be significantly improved with a large enough building stock sample. To apply this reasoning more generally, building stock energy modelers should be aware of the potential deviation due to stock size for lesser-represented portions of the building stock, particularly if they are relevant to the analysis they are pursuing.

## 6.6 Case study

The government of Québec has the objective of reducing greenhouse gas emissions (GHG) related to space heating in buildings by 50% for 2030 (Government of Québec 2020). While the proposed building stock model is not a long-term energy projection model, a comparative assessment between two or more stock configurations can be performed. In order to study the effect of energy consumption changes on GHG emissions, the emissions factors for the province of Québec are first

presented. The proposed scenarios are then compared, illustrating the impact of changes to the single-family dwelling space heating market on the energy consumption, GHG emissions and peak electricity use.

### 6.6.1 Greenhouse gas emission factors

As described previously, energy use in single-family dwellings in the province of Québec, Canada, consists primarily of electricity, natural gas, heating oil and wood. These energy sources each have distinct CO<sub>2</sub> equivalent emission factors, which are provided in Table 6.9 for electricity, natural gas, heating oil and wood.

Table 6.9 Greenhouse gas emission factors for energy sources in the province of Québec

Energy source	gCO <sub>2</sub> eq•kWh <sup>-1</sup>	Ref.
Electricity	2.0	TÉQ (2019)
Natural gas	178.3	TÉQ (2019)
Heating oil	254.9	TÉQ (2019)
Wood	84.5	NRCan (2017b)

The emissions factors in Table 6.9 are considered constant for the entire year, with the exception of electricity. Electricity production in the province of Québec is predominantly hydroelectric and is estimated to generate CO<sub>2</sub> equivalent emissions at the rate of 2.0 gCO<sub>2</sub>eq•kWh<sup>-1</sup> (Transition Énergétique Québec 2019). During peak electricity usage hours, a non-negligible portion of the electricity in the province is imported from neighbouring provinces and states, which is called *short-term imported electricity* and is illustrated in red in Figure 6.9 for December 2017 (Régie de l'énergie 2017).

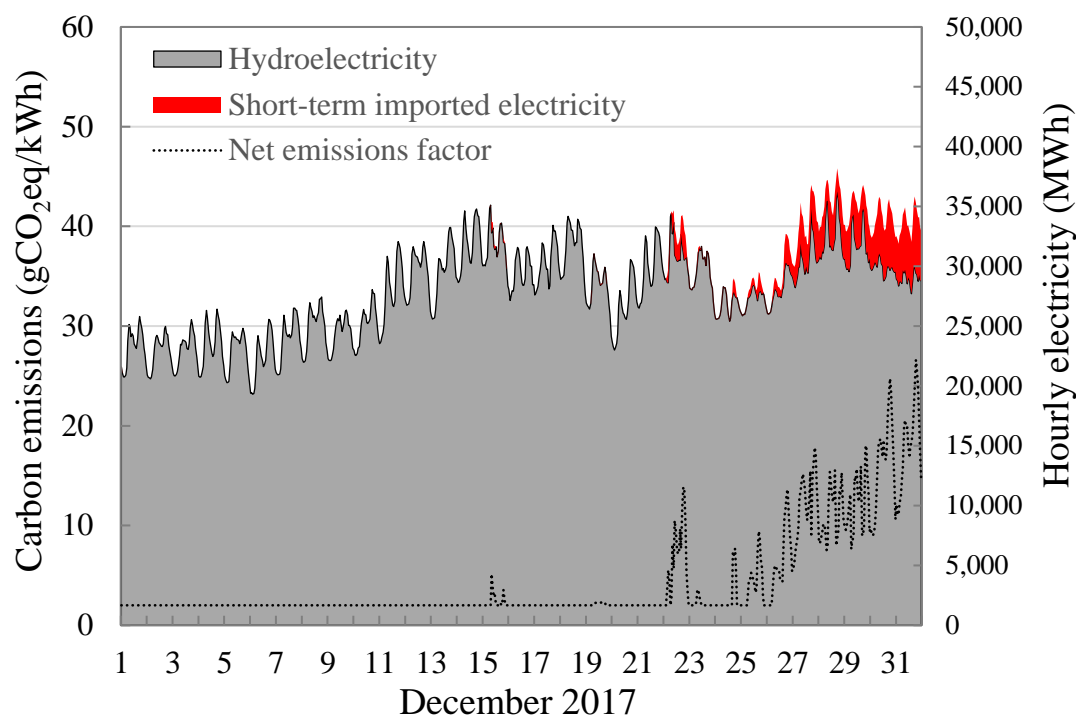


Figure 6.9 Calculated CO<sub>2</sub> equivalent emission rates in December 2017 for electricity in the province of Québec and electricity usage by source for all sectors

According to the local electricity distributor, 97.5% of short term electricity purchases are from Ontario (Hydro-Québec 2017). To simplify the GHG case study analysis, the short term electricity imports for Québec are calculated using marginal seasonal emission factors for the province of Ontario, which vary hourly between 90 and 150 gCO<sub>2</sub>eq•kWh<sup>-1</sup> (The Atmospheric Fund 2019). Combining the base electricity emissions factor of 2.0 gCO<sub>2</sub>eq•kWh<sup>-1</sup> for local hydroelectricity with the short term imported electricity from Ontario results in a net emissions factor that varies approximately between 2.0 and 26.0 gCO<sub>2</sub>eq•kWh<sup>-1</sup>, depending on the day and time of year (Figure 6.9).

While a large increase to the peak electricity usage would theoretically result in more short-term imports, in reality sweeping changes to the building stock would not occur in a short time frame. It is likely that the local electricity distributor would adjust the hydroelectric production to account

for any large increases to the stock electricity use. Therefore, emissions rates for additional electricity use are calculated at the rates as the current stock, as described above.

### **6.6.2 Case study: 50% reduction in GHG emissions for space heating**

In order to study the impact of reducing GHG emissions due to space heating, three scenarios are compared:

1. Base case: the status quo single-family building stock for the province of Québec, Canada.
2. Scenario 1: 50% of heating systems with non-electric energy sources are converted to baseboard electric heating (heating system H3, Table 6.3), a common cheap alternative widely used in the province of Québec.
3. Scenario 2: 50% of heating systems with non-electric energy sources are converted to cold climate heat pump heating systems.

Cases are compared based on the annual space heating energy consumption, the GHG emissions for each energy source and the maximum peak load. Heating system distributions are updated by modifying the probabilities presented in Table 6.5, shifting heating systems with non-electric energy sources to electric systems as described in the scenario descriptions. In the case of Scenario 2, the coefficient of performance (COP) is implemented as a function of outdoor air temperature based on measured data (R. K. Johnson 2013).

The total stock emissions and space heating energy consumption are presented for the base case and two scenarios in Figure 6.10. Total emissions and energy consumption are similar to the reference data by NRCan (2017b). Most of the emissions in the province originate from the non-electric energy sources, and therefore reducing the heating systems using those energy types by 50% has the desired effect of reducing overall emissions by nearly 50%. In terms of the energy consumption for Scenarios 1 and 2, reductions in total space heating energy use for single-family dwellings reach 10% and 21%, respectively. These decreases in space heating energy are largely due to the improvements in heating system performance when comparing electric systems to their nonelectric counterparts.

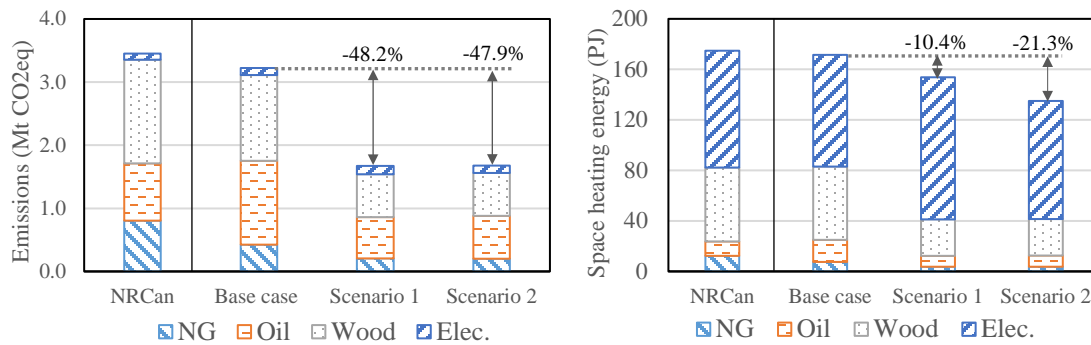


Figure 6.10 Annual CO<sub>2</sub> equivalent emissions and total space heating energy consumption for the studied cases. Stock reference data from Natural Resources Canada is also provided for comparison (NRCan 2017)

The difference in peak loads between the base case and Scenarios 1 and 2 are compared in Figure 6.11, which are illustrated for the month of December 2017. While emissions and annual space heating energy are reduced for both scenarios, the peak electricity use increases significantly. For Scenario 1, shifting nonelectric heating to baseboard heaters has the result of increasing the peak load over the normal SFD stock value by approximately 35%, or the equivalent of 4000 MW. For Scenario 2, the improved efficiency of the heat pumps mitigate the impact on the peak load, but still increase it by up to 20%, or approximately 2500 MW. During milder periods, the electricity load can occasionally decrease below the normal level of the base case scenario due to the higher efficiency of the heat pumps at milder temperatures. However, decreasing the load at other moments of the year does not aid the electricity distributor, as the maximum electricity production capacity of the stock is sized for the peak usage of the province, which would increase under both scenarios.

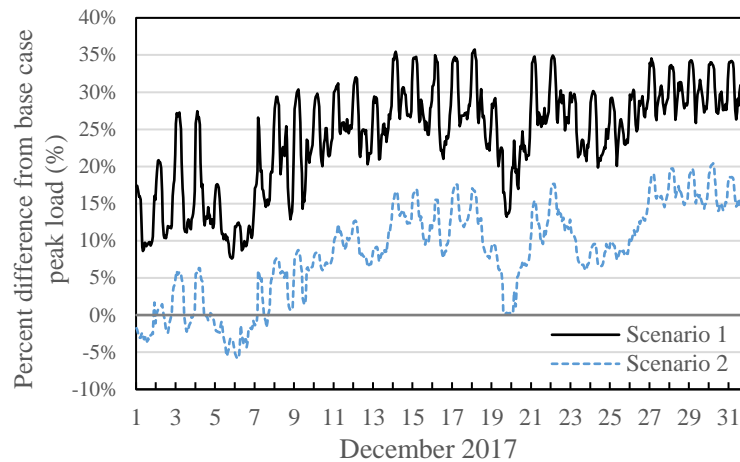


Figure 6.11 Peak load percent difference for Scenarios 1 and 2 with respect to the base case scenario

The case study illustrates how the QSFBSSEM can be used to compare greenhouse gas scenarios for different heating system distributions across the province. Similar studies could be performed at the regional level or considering other aspects of the building stock, such as retrofitting the building envelope or installing cooling systems. The case study demonstrates that the studied building stock can achieve the desired greenhouse gas reductions by shifting heating systems to electric alternatives, but at the cost of a significant increase to the peak electricity load. If not addressed properly by the local electricity distributor, the marginal electricity use for such scenarios could result in additional short-term electricity purchases from neighbouring sources with much higher electricity emission rates.

## 6.7 Conclusion

The bottom-up white-box building stock energy model developed by the authors represents the single-family dwelling market for the province of Québec, Canada. The characterization process applied to the provincial stock data ensures that each region reflects the real distribution of systems and building characteristics according to the best available information. By implementing region-specific probability distributions, the stock model can then be applied to different areas of the

province, such as the city of Montreal, rural areas, or to the entire province. The model is described using the stock energy model labelling system proposed by Langevin et al. (2020):

**Country:** Canada (province of Québec)

**Model name:** QSFBSSEM

**Model use:** A static bottom-up white-box stock model for comparative assessment of residential stock energy use. Energy can be categorized by end-use and by source, and peak electricity demand can be assessed. Appropriate for technological evaluation and greenhouse gas emissions studies.

**Model classification quadrant:** Q4 (physics simulation)

**Additional details:** N/A.

The energy consumption for various end-uses and energy source totals is compared between the model and reference data for attached, detached, and all single-family dwellings. In total, 30 points of comparison, aggregated annually and geographically, are verified to ensure the model represents all facets of the studied building stock, including the space heating, water heating, lighting, appliances, space cooling, etc. The model provides an accurate prediction of the stock energy consumption, though some non-electric energy sources are under- or over-predicted. The total stock energy consumption of the model is within 1.6% of the provincial single-family dwelling stock data for 2017.

Technological changes can be evaluated using the building stock model, which introduces the possibility of comparative assessments for a variety of applications. The capability to introduce technological changes to the building stock and evaluate them addresses one of the issues raised by Booth et al. (2012), described as the *flexibility* of building stock models. One of the key features of white-box models is the capability to evaluate technological changes immediately with simulation, rather than waiting for data to develop a statistics-based model. For the proposed model, the probability distributions can be altered to perform a direct comparison between different configurations of building stock characteristics, or a new technology can be introduced to a portion of the stock.

The case study presented in the paper illustrates how a provincial greenhouse gas reduction target can potentially be achieved, but at the cost of a significant increase in peak electricity load. The proposed building stock energy model allows for an hourly comparison of electricity usage of the building stock, which allows for additional time-dependent analysis of greenhouse gas emissions

and evaluation of the best measures based on marginal emission rates. Future studies can leverage the flexibility of the proposed model to evaluate a wide variety of building stock configurations.

## 6.8 References

- Booth, A.T., R. Choudhary, and D.J. Spiegelhalter. 2012. “Handling Uncertainty in Housing Stock Models.” *Building and Environment* 48 (February). Pergamon: 35–47.  
doi:10.1016/J.BUILDENV.2011.08.016.
- Government of Canada. 2021. “Engineering Climate Datasets.” Engineering Climate Services.  
[https://climate.weather.gc.ca/prods\\_servs/engineering\\_e.html](https://climate.weather.gc.ca/prods_servs/engineering_e.html).
- Government of Québec. 2020. 2030 Plan for a Green Economy. Québec, Canada.
- Hong, Tianzhen, Yixing Chen, Sang Hoon Lee, and Mary Ann Piette. 2016. “CityBES: A Web-Based Platform to Support City-Scale Building Energy Efficiency.”  
doi:10.1145/12345.67890.
- Hydro-Québec. 2017. Sustainable Development Plan 2020-2024.
- Hydro-Québec. 2021. “Rate DT – Dual Energy | Hydro-Québec.”  
<https://www.hydroquebec.com/residential/customer-space/rates/rate-dt.html>.
- Johnson, R K. 2013. Measured Performance of a Low Temperature Air Source Heat Pump Consortium for Advanced Residential Buildings. Oak Ridge, TN, USA.
- Kasten, Fritz, and Gerhard Czeplak. 1980. “Solar and Terrestrial Radiation Dependent on the Amount and Type of Cloud.” *Solar Energy* 24 (2). Pergamon: 177–189. doi:10.1016/0038-092X(80)90391-6.
- Kavgic, M., A. Mavrogianni, D. Mumovic, A. Summerfield, Z. Stevanovic, and M. Djurovic-Petrovic. 2010. “A Review of Bottom-up Building Stock Models for Energy Consumption in the Residential Sector.” *Building and Environment* 45 (7): 1683–1697.  
doi:10.1016/j.buildenv.2010.01.021.



- Klein, S.A., W.A. Beckman, J.W. Mitchell, J.A. Duffie, N.A. Duffie, T.L. Freeman, J.C. Mitchell, et al. 2017. “TRNSYS 18: A Transient System Simulation Program, Solar Energy Laboratory, University of Wisconsin, Madison, USA.” <http://sel.me.wisc.edu/trnsys>.
- Langevin, J., J. L. Reyna, S. Ebrahimigharehbaghi, N. Sandberg, P. Fennell, C. Nägeli, J. Laverge, et al. 2020. “Developing a Common Approach for Classifying Building Stock Energy Models.” *Renewable and Sustainable Energy Reviews* 133 (November). Pergamon: 110276. doi:10.1016/J.RSER.2020.110276.
- Mathworks Inc. 2018. “Matlab R2018b.”
- McKenna, Eoghan, and Murray Thomson. 2016. “High-Resolution Stochastic Integrated Thermal–Electrical Domestic Demand Model.” *Applied Energy* 165: 445–461.
- Nägeli, Claudio, Clara Camarasa, Martin Jakob, Giacomo Catenazzi, and York Ostermeyer. 2018. “Synthetic Building Stocks as a Way to Assess the Energy Demand and Greenhouse Gas Emissions of National Building Stocks.” *Energy and Buildings* 173 (August). Elsevier: 443–460. doi:10.1016/J.ENBUILD.2018.05.055.
- National Renewable Energy Laboratory. 2021. “National Baseline Data Viewer - ResStock.” [https://resstock.nrel.gov/dataviewer/efs\\_v2\\_base](https://resstock.nrel.gov/dataviewer/efs_v2_base).
- Neale, Adam, Michaël Kummert, and Michel Bernier. 2020. “Development of a Stochastic Virtual Smart Meter Data Set for a Residential Building Stock – Methodology and Sample Data.” *Journal of Building Performance Simulation* 13 (5): 583–605. doi:10.1080/19401493.2020.1800096.
- New, Joshua Ryan, Mark B. Adams, Piljae Im, Hsiuhan Lexie Yang, Joshua C. Hambrick, William E. Copeland, Lilian B. Bruce, and James A. Ingraham. 2018. “Automatic Building Energy Model Creation (AutoBEM) for Urban-Scale Energy Modeling and Assessment of Value Propositions for Electric Utilities.” In *International Conference on Energy Engineering and Smart Grids*, 5. Oak Ridge, TN, USA: USDOE Office of Energy Efficiency and Renewable Energy, USDOE Office of Electricity.

- NRCan (Natural Resources Canada). 2014. “Canada’s Secondary Energy Use by Sector, End Use and Subsector.” Energy Use Data Handbook Tables.
- NRCan (Natural Resources Canada). 2015. “2015 Survey of Household Energy Use Data Tables by Census Metropolitan Area.” National Energy Use Database.  
<https://oee.nrcan.gc.ca/corporate/statistics/neud/dpa/menus/sheu-cma/2015/tables.cfm>.
- NRCan (Natural Resources Canada). 2017. “National Energy Use Database.” Natural Resources Canada.
- NRCan (Natural Resources Canada). 2018. “Energuide Housing Database.”
- NYSERDA. 2013. Multifamily Performance Program Gas Furnace Electricity Usage Background.
- Oak Ridge National Laboratories. 2021. “Virtual EPB.”  
[https://evenstar.ornl.gov/autobem/virtual\\_epb/](https://evenstar.ornl.gov/autobem/virtual_epb/).
- Régie de l’énergie. 2017. “Relevés Des Livraisons d’énergie En Vertu de l’entente Globale Cadre Pour La Période Du 1er Janvier Au 31 Décembre 2017 - Version Amendée .” Audiences et Décisions, D-2016-143. [http://www.regie-energie.qc.ca/audiences/Suivis/Suivi\\_HQD\\_D-2016-143.html](http://www.regie-energie.qc.ca/audiences/Suivis/Suivi_HQD_D-2016-143.html).
- Reinhart, Christoph F., and Carlos Cerezo Davila. 2016. “Urban Building Energy Modeling - A Review of a Nascent Field.” Building and Environment 97 (February). Elsevier Ltd: 196–202. doi:10.1016/j.buildenv.2015.12.001.
- Reinhart, Christoph F, Timur Dogan, Alstan Jakubiec, Tarek Rakha, and Andrew Sang. 2013. “UMI - an Urban Simulation Environment for Building Energy Use, Daylighting and Walkability.” In 13th Conference of International Building Performance Association, 476–483. Chambéry, France.
- Sherman, M.H., and D.T. Grimsrud. 1980. “Infiltration-Pressurization Correlation: Simplified Physical Modeling.” ASHRAE Transactions 86 (2): 778.

- Sokol, Julia, Carlos Cerezo Davila, Christoph F. Reinhart, C. Cerezo, and Christoph F. Reinhart. 2016. "Validation of a Bayesian-Based Method for Defining Residential Archetypes in Urban Building Energy Models." *Energy and Buildings* 134. Elsevier B.V.: 11–24.
- Statistics Canada. 2016. "2016 Census." Statistics Canada. <http://www12.statcan.gc.ca/census-recensement/2016/dp-pd/index-eng.cfm>.
- Statistics Canada. 2019. "Annual Demographic Estimates: Subprovincial Areas." <https://www150.statcan.gc.ca/n1/pub/91-214-x/2020001/section01-eng.htm>.
- Swan, Lukas G., and V. Ismet Ugursal. 2009. "Modeling of End-Use Energy Consumption in the Residential Sector: A Review of Modeling Techniques." *Renewable and Sustainable Energy Reviews*. doi:10.1016/j.rser.2008.09.033.
- Taniguchi-Matsuoka, Ayako, Yoshiyuki Shimoda, Minami Sugiyama, Yusuke Kurokawa, Haruka Matoba, Tomoya Yamasaki, Taro Morikuni, and Yohei Yamaguchi. 2020. "Evaluating Japan's National Greenhouse Gas Reduction Policy Using a Bottom-up Residential End-Use Energy Simulation Model." *Applied Energy* 279 (December). Elsevier: 115792. doi:10.1016/J.APENERGY.2020.115792.
- The Atmospheric Fund. 2019. *A Clearer View on Ontario's Emissions Electricity Emissions Factors and Guidelines*.
- Transition Énergétique Québec. 2019. "Emission Factors and Conversion." <https://transitionenergetique.gouv.qc.ca/fileadmin/medias/pdf/FacteursEmission.pdf>.
- Ugursal, V I. 2017. *Canadian Building Stock Data: Notes from the Stock Modeling Workshop*. Ottawa, Canada.
- US DOE. 2013. *EnergyPlus Engineering Reference - The Reference to EnergyPlus Calculations* (v 8.1). Washington, DC, USA: US Department of Energy.
- Wilson, Eric, Craig Christensen, Scott Horowitz, Joseph Robertson, and Jeff Maguire. 2017. *Energy Efficiency Potential in the U.S. Single-Family Housing Stock*. Denver, CO, USA.

## Appendix 6.1: Weather stations for each region of the studied building stock

Table 6.10 Weather stations for each region of the studied building stock. CMA: census metropolitan area, CA: census agglomeration

Number	Region type	Fraction	Location name	Nearest weather station
R1-01	CA	0.0156	Alma	CAN_QC_JONQUIERE_7063370
R1-02	CA	0.0144	Baie-Comeau	CAN_QC_BAIE-COMEAU_704S001
R1-03	CA	0.0019	Campbellton	CAN_NB_CHARLO-AUTO_8100885
R1-04	CA	0.0053	Cowansville	CAN_QC_FRELIGHSBURG_7022579
R1-05	CA	0.0082	Dolbeau-Mistassini	CAN_QC_NORMANDIN_7065639
R1-06	CA	0.0440	Drummondville	CAN_QC_NICOLET_7025442
R1-07	CA	0.0362	Granby	CAN_QC_FRELIGHSBURG_7022579
R1-08	CA	0.0057	Hawkesbury	CAN_QC_MONTREAL-MIRABEL-INTL-A_7034900
R1-09	CA	0.0207	Joliette	CAN_QC_L'ASSOMPTION_7014160
R1-10	CA	0.0058	Lachute	CAN_QC_MONTREAL-MIRABEL-INTL-A_7034900
R1-11	CA	0.0104	Matane	CAN_QC_AMQUI_7050145
R1-12	CA	0.0288	Rimouski	CAN_QC_POINTE-AUX-PERE-(INRS)_7056068
R1-13	CA	0.0142	Rivière-du-Loup	CAN_QC_RIVIERE-DU-LOUP_7056616
R1-14	CA	0.0199	Rouyn-Noranda	CAN_QC_ROUYN-NORANDA-A_7086719
R1-15	CA	0.0175	Saint-Georges	CAN_QC_BEAUCEVILLE_7028754
R1-16	CA	0.0217	Saint-Hyacinthe	CAN_QC_MONTREAL-ST-HUBERT_7027329
R1-17	CA	0.0073	Sainte-Marie	CAN_QC_BEAUCEVILLE_7028754
R1-18	CA	0.0181	Salaberry-de-Valleyfield	CAN_QC_ST-ANICET-1_702FQLF
R1-19	CA	0.0125	Sept-Îles	CAN_QC_SEPT-ILES-A_7047911
R1-20	CA	0.0261	Shawinigan	CAN_QC_SHAWINIGAN_7018001
R1-21	CA	0.0220	Sorel-Tracy	CAN_QC_LAC-SAINT-PIERRE_701LP0N
R1-22	CA	0.0160	Thetford Mines	CAN_QC_BEAUCEVILLE_7028754
R1-23	CA	0.0151	Val-d'Or	CAN_QC_VAL-D'OR_7098603
R1-24	CA	0.0235	Victoriaville	CAN_QC_LEMIEUX_701Q009
R1	Québec non-CMA	0.4108		
R2	CMA	0.0233	Saguenay	CAN_QC_JONQUIERE_7063370
R3	CMA	0.0949	Québec	CAN_QC_QUEBEC-INTL-A_7016293
R4	CMA	0.0262	Sherbrooke	CAN_QC_LENNOXVILLE_7024280
R5	CMA	0.0214	Trois-Rivières	CAN_QC_NICOLET_7025442
R6	CMA	0.3756	Montréal	CAN_QC_MONTREAL-INTL-A_7025251
R7	CMA	0.0478	Ottawa - Gatineau	CAN_ON_OTTAWA-INTL-A_6106001

## Appendix 6.2: Total and opaque cloud cover data correction

The cloud fraction can be calculated with Equation (6.2) using the global and diffuse horizontal radiation values that are available in the CWEEDs weather data (Kasten and Czeplak 1980).

$$f_{cloud} = 10 * \left( 1.4286 \frac{E_{dif}}{E_{glob,h}} - 0.3 \right)^{0.5} \quad (6.2)$$

where  $E_{dif}$  and  $E_{glob,h}$  are the diffuse and global horizontal radiation (Wh/m<sup>2</sup>), respectively, and  $f_{cloud}$  is the cloud fraction in tenths. The Total and Opaque Sky Cover values in the CWEEDs files are assumed equal to the cloud fraction and are rounded to the nearest whole integer, which is common practice (Government of Canada 2021). Since the global horizontal radiation in CWEEDs data is zero at night, the cloud cover data during night time is linearly interpolated between the last available value for cloud cover and the first available data point the next morning, as illustrated in Table 6.11.

In summary, the general approach for completing the cloud cover data is as follows:

1. Verify whether any monitored data for Total Sky Cover was available, usually at 3 hour intervals. If yes, set the values for Total Sky Cover for hours 2 and 3 equal to the first hour and repeat for the whole year.
2. If no measured data is available, use Equation (6.2) to determine the cloud fraction for the hours where the global and diffuse horizontal radiation is available, and use linear interpolation to complete the night time values.

In all cases, the Opaque Sky Cover is assumed equal to the Total Sky Cover.

Table 6.11 Example cloud cover data generated and filled for a 24 hour period. Night time values highlighted in grey are filled using linear interpolation between the two data points in bold text

Hour of the day	Global horizontal radiation (Wh/m <sup>2</sup> )	Diffuse horizontal radiation (Wh/m <sup>2</sup> )	Cloud cover without interpolation (tenths)	Cloud cover with interpolation (tenths)
13	827	276	4	4
14	889	254	3	3
15	249	220	10	10
16	654	189	3	3
17	468	93	0	0
18	305	59	0	0
19	135	43	4	4
20	21	14	<b>8</b>	<b>8</b>
21	0	0	0	8
22	0	0	0	7
23	0	0	0	7
24	0	0	0	7
1	0	0	0	6
2	0	0	0	6
3	0	0	0	6
4	0	0	0	5
5	0	0	0	5
6	142	51	<b>5</b>	<b>5</b>
7	310	67	1	1
8	487	83	0	0
9	647	90	0	0
10	780	212	3	3
11	225	216	10	10
12	229	227	10	10

### Appendix 6.3: QSFBSM overview

A parallel processing approach is adopted to distribute the task to multiple workers due to the number of houses required for a building stock simulation. Typically 20 workers are assigned to the task of modeling the building stock. As a point of reference, each set of 1000 houses modeled using the stock model requires approximately 30 minutes on a server with an Intel Core i9-7920X processor @2.9 GHz, 128 GB of RAM @2133 MHz and a SATA III solid-state hard drive. The modeling process is divided into a number of steps (Figure 6.12):

1. **Model initialization:** a new house simulation is initialized. Conditions for the stock simulation are registered at this stage. For example, limiting the stock simulation to a single region instead of all seven regions of the province.
2. **Input generation:** the building parameters (Figure 6.13) are generated according to the probability distributions from the stock characterisation process and based on the scope of the simulation specified in (1).
3. **TRNSYS file preparation:** required TRNSYS simulation files are automatically prepared for the current building energy simulation according to the randomly generated building characteristics from step (2).
4. **TRNSYS simulation:** an annual building energy simulation is performed and energy consumption values are output at 15-minute intervals. Input characteristics for each house are also retained and matched with the energy consumption.
5. **Worker data:** when the desired number of house simulations is reached, the worker data is saved.

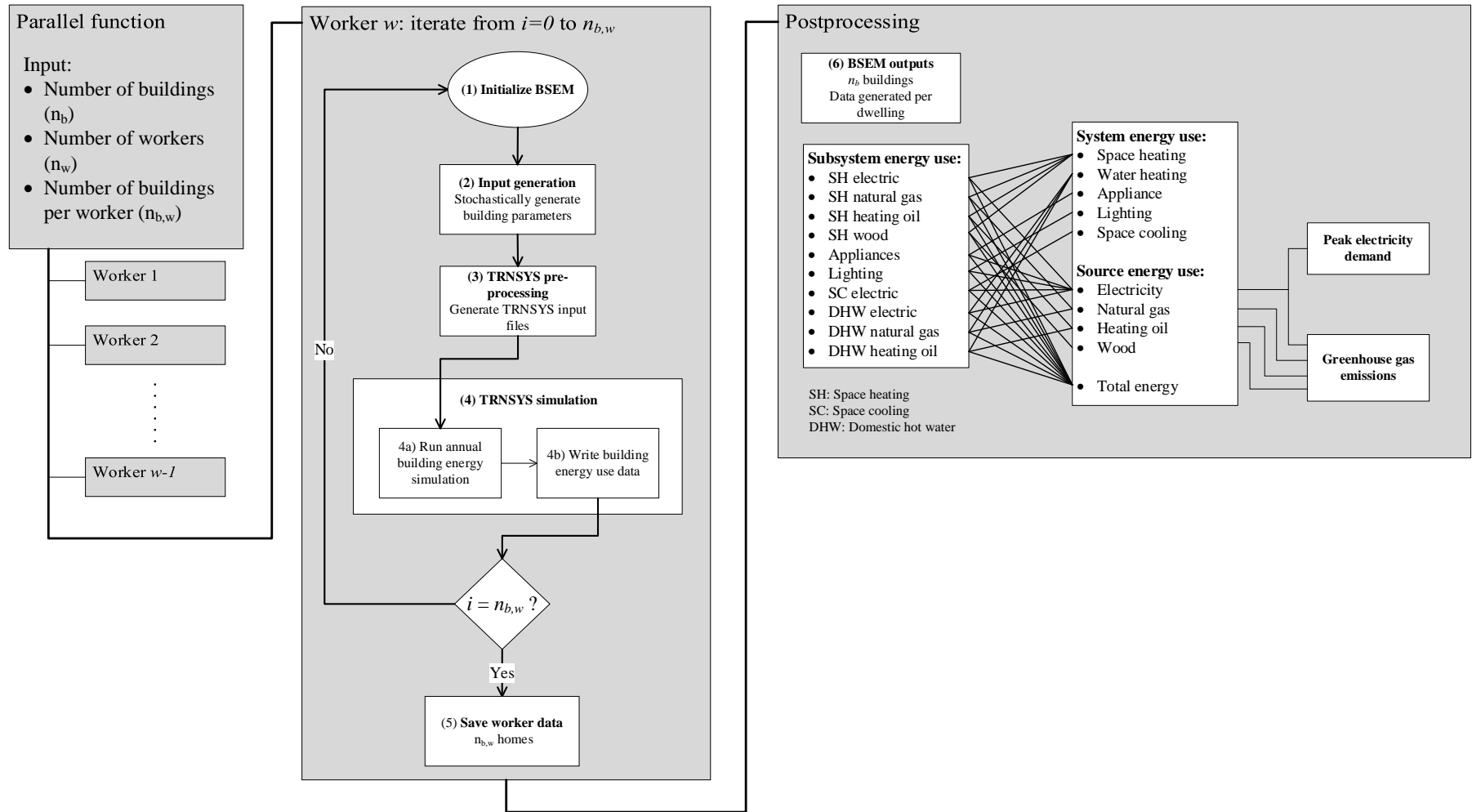


Figure 6.12 General building stock energy modeling approach and model outputs



6. **Postprocessing:** worker data are compiled into a single data set. Data is categorized by end-use and energy source. Total stock electricity data is retained at 15-minute intervals while natural gas, heating oil and wood energy consumption is saved annually.

The Matlab software is used as a platform to distribute the building energy simulations to individual workers, generate the building input files and launch the TRNSYS building energy simulations (Mathworks Inc. 2018a). The building parameters and their interdependencies are illustrated in Figure 6.13. The result is a flexible building stock energy model that can generate virtually any number of homes according to the input parameters provided to the model.

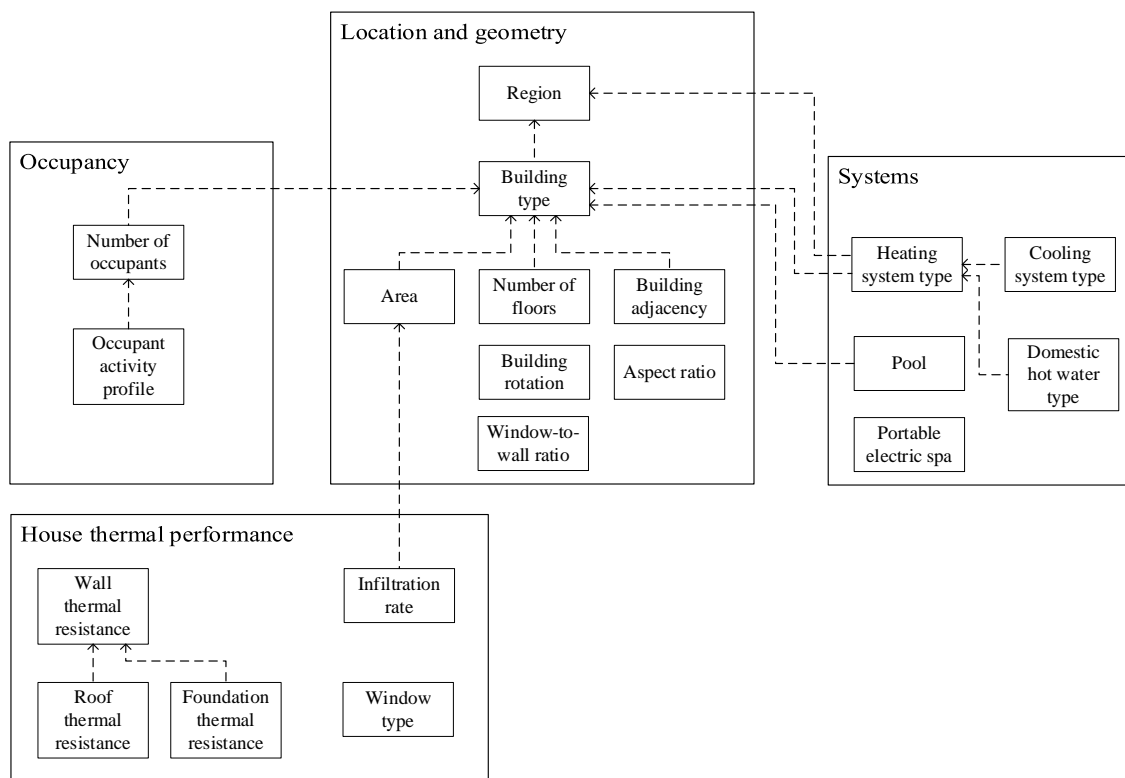


Figure 6.13 Building parameters generated for each house. Dependencies are illustrated with dashed arrows

As mentioned previously, the modeling principles in TRNSYS and the approach for specific aspects of the building energy simulation, such as the internal loads, were described in another publication (Neale, Kummert, and Bernier 2020a). Some features are described in brief detail in Table 6.12 for clarity, and interested readers can refer to the original publication for more detail.

Table 6.12 Overview of modeling choices for the building energy simulation of each house

Model aspect	Detail	Reference (if applicable)
Geometry	Each house is modeled as a single-zone dwelling with a finished basement. Wall surfaces shared with adjacent dwellings, such as in row houses, are considered adiabatic. The window-to-wall ratio is applied on all aboveground wall surfaces not shared by another dwelling.	
Thermal performance characteristics	Wall, roof and foundation thermal resistance are applied separately to their corresponding surfaces. Infiltration rate is calculated using the Sherman-Grimsrud method. Window type is selected from a list of 24 single-, double-, and triple-glazed models.	Sherman and Grimsrud (1980)
Internal loads	Internal loads due to occupancy are produced using the CREST stochastic occupancy model, which includes the lighting and appliance loads. Loads are applied as fractional convective and radiative heat gains to the indoor environment.	McKenna and Thomson (2016)
Systems	Domestic hot water energy demand profiles are generated separately from the stock simulation using the TRNSYS software. Hot water use is determined using the CREST stochastic occupancy model as a function of the number of occupants currently active in the home.  Heating systems are assumed to have infinite capacity unless otherwise specified in Table 6.6. Other characteristics of the heating systems are applied to address the building heating load as described in Table 6.6.  Cooling systems are assumed to have infinite capacity.	McKenna and Thomson (2016)

## **CHAPTER 7      GENERAL DISCUSSION**

The literature review identified a variety of current needs and limitations related to building stock energy modeling, and this thesis attempts to address some of those aspects. Ugursal (2017) describes a clear industry need for more stock energy prediction tools in Canada expressed during a workshop on building stock modeling held in 2017, where stakeholders agreed that the tools available at the time were insufficient for industry, academic and government purposes. In addition, the literature identifies specific issues that have yet to be resolved in building stock energy modeling. These include limitations such as the lack of data available to characterize a building stock, the accuracy of building stock models, limited computational resources, and the capability of models to adapt to new technological changes (Booth, Choudhary, and Spiegelhalter 2012; J. Langevin et al. 2020). The main contributions of this thesis are described in the following section.

### **7.1 Review of journal paper contributions**

The literature review presented in Chapter 1, and in the journal paper in Chapter 4, has shown that there are no publicly available smart meter data sets detailed enough to perform a classification study on residential smart meter data. The first objective of this thesis aims to solve this issue by developing a virtual smart meter (VSM) data set suitable for classification studies, with the goal to illustrate the possibility to apply the technique to building stock model characterization.

The journal article in Chapter 4 presents a general methodology to develop a VSM data set, which is a set of simulated single-family dwelling electricity consumption profiles with known building parameters. The methodology is generalized so that other interested researchers can develop their own VSM data profiles with their proper stock characterization data. An open-source sample data set is included with the paper and is available online for anyone to download (Neale, Kummert, and Bernier 2020b). The data set is designed for classification studies and is unique in that no other smart meter data set has the same quantity of building data with which to develop predictive models.

The development of the virtual smart meter data set allowed for the study of classification techniques, with the objective of predicting building characteristics from smart meter data, whether virtual or measured. The second journal article, presented in Chapter 5, performs a detailed supervised machine learning classification study on the virtual smart meter data set developed in

Chapter 4. Few relevant classification studies were found in the literature due to the lack of appropriate data sets, and none tested classification on as many different building parameters. The results of the paper provide a clear understanding on the impact of the features (i.e. the electricity use at different frequencies) and number of buildings on the classification accuracy. The size of the smart meter data set must exceed the number of features by a significant margin to ensure stable classification accuracy results, which can instruct future development efforts for smart meter monitoring campaigns. The article provides guidelines on designing a real smart meter data set with known building parameters for the purpose of machine learning classification.

The third journal article, presented in Chapter 6, provides the framework and description for a new bottom-up white-box building stock energy model for single-family buildings in the province of Québec, Canada. The model is called the Québec Single-Family Building Stock Energy Model (QSFBSM). The details of the characterization process are provided such that other researchers can follow the same procedure for their own building stock. The accuracy of the model is evaluated based on 30 different energy totals, including by building type, end-use, and energy source, with generally good agreement. The impact of the sample size of dwellings produced by the stock model is discussed, with guidelines on the expected deviation of the energy consumption result based on the sample size. In addition, a case study presented in the journal paper illustrates the usefulness of a stock model able to calculate time-dependent energy consumption values for peak load determination. Most stock models in the literature are unable to evaluate the peak load of a building stock, which is a critical concern in the province of Québec with the current electrification trend of the provincial government (Government of Québec 2020).

A more general contribution related to the development of the VSM data set in journal article 1 and the building stock energy model in journal article 3 is the detailed characterization of the building parameters used to produce the probability distributions. The process required an extensive data gathering effort and statistical analysis to establish the dependence between the parameters and apply Bayes Theorem to create dependent probability distributions based on prior data. These aspects are described in part in Chapters 4 and 6.

## 7.2 Other contributions

In addition to the contributions in the journal articles, an electricity smart meter comparison tool was developed to visualize differences between different building types and their resulting electricity consumption profiles, illustrated in Figure 7.1. While still in development, the visualization tool provided a valuable asset used to qualitatively and quantitatively compare the electricity consumption profiles for different types of single-family dwellings.

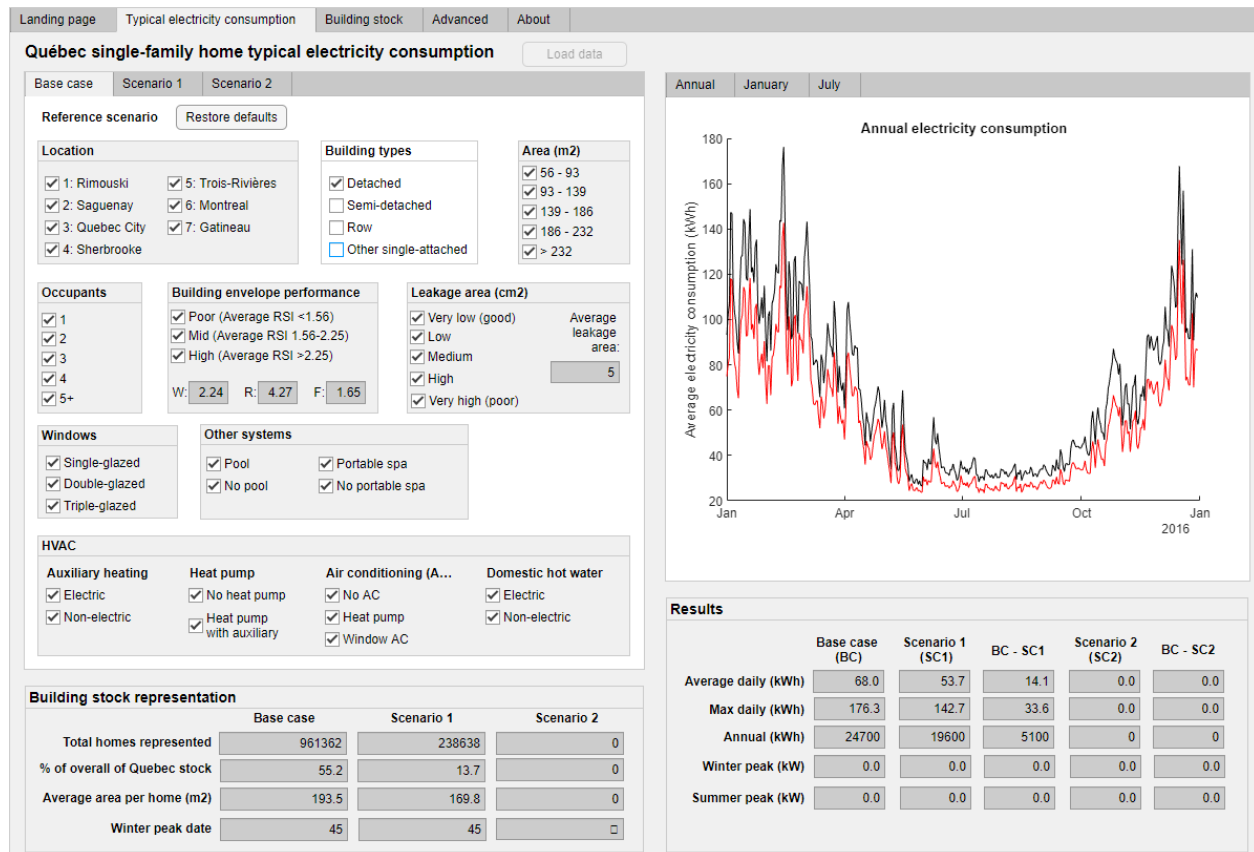


Figure 7.1 Electricity consumption profile comparison tool

## **CHAPTER 8      CONCLUSION AND RECOMMENDATIONS**

The contributions of this thesis focus on improving a number of limitations related to building stock energy modeling. Foremost among these factors is the lack of available tools for the prediction of stock energy consumption and peak load prediction, which are becoming increasingly necessary for a variety of applications. While much work has been done in recent years in the field of stock modeling, other limitations exist, such as the lack of data for the segmentation and characterization processes commonly used for developing stock models.

This thesis presents three main steps that were taken to address the lack of data for stock characterization and to develop a new single-family dwelling building stock energy model for the province of Québec.

A virtual smart meter data set with a variety of known building parameters is developed. The data set is produced due to the lack of public measured electricity data sets for residential buildings that include sufficient information about the dwellings. The development of the data set required the characterization of a variety of building parameters to allow for the modeled dwellings to correspond as closely as possible to real buildings in the stock. Designing and producing a VSM data set where each building has a set of known building characteristics introduces the possibility of evaluating supervised machine learning classification as a building stock characterization technique.

Linear discriminant analysis (LDA) is used to develop predictive models that create relationships between the electricity consumption of a house and each building characteristic, such as the heated surface area or the number of occupants. The classification accuracy of the predictive models demonstrates that LDA is effective at identifying building characteristics within a certain range of values, and generally much more accurate than randomly guessing the outcome. The classification results indicate that a predictive model could read any smart meter data and provide a reasonable approximation of certain building characteristics. This illustrates that smart meter data could be an interesting source of data for building stock characterization.

Finally, a new single-family dwelling (SFD) building stock energy model for the province of Québec is developed (called the QSFBSEM). The model predicts the total SFD energy consumption within 1.6% of reference data and has very good agreement across a variety of energy

end-uses, and energy source totals and dwelling subtypes, which is significantly better than the accuracy of building archetypes identified in the literature. The potential of the QSFBSSEM is demonstrated in a case study predicting the impact of heating system changes on the stock greenhouse gas emissions, total energy consumption and peak electricity load. While government targeted emissions reductions are possible to achieve by reducing non-electric heating systems, it is at the cost of a substantial increase to the peak electricity usage of the single-family dwelling building stock, by up to 4000 MW (nearly 10% of the total provincial peak electricity usage) during peak heating periods.

## 8.1 Recommendations

This thesis provides a new technique for characterizing a building stock using smart meter data, which applies supervised machine learning classification to predict building parameters (such as wall thermal resistance for example) from electricity consumption values. While the method is shown to have promising results by predicting a variety of building parameters with significantly higher accuracy than randomly assigning values, it could be further improved by applying it to a set of real smart meter data with known building parameters. A future measurement campaign would benefit from the guidelines provided in the results of the classification study, in terms of the number of buildings required to be able to develop accurate predictive models and the parameters that should be recorded for each building in the data set. Other classification algorithms should be evaluated and compared to linear discriminant analysis, as some building parameters may be better classified with different classification techniques.

The QSFBSSEM developed for the Québec single-family dwelling building stock accurately predicts the energy consumption across a variety of end-use and energy source categories. The capability of the stock model to determine energy consumption at hourly and subhourly intervals allows for a wide range of applications, such as the evaluation of energy conservation measures targeting specific building technologies, or the quantification of greenhouse gas emissions savings due to a particular measure. The QSFBSSEM could be improved in a number of ways. First, an interface is recommended to facilitate launching simulations and to modify the characteristics of the building stock for future case studies using the model. In addition, a next logical step would be

to add multi-residential unit buildings to the stock model in order to encompass the entire residential building stock of the province. Refining the internal loads of the residential buildings to be more specific to the province would be an additional improvement, as the external load tool is not specifically made for Canadian occupants, which could result in some cultural differences in domestic load usage. Finally, *decision-making* is one of the key limitations of stock model development identified by Booth et al. (2012), which states that the impact of uncertainty on the outcome of stock modeling should be evaluated. It would be relevant in the future to evaluate this aspect for the QSBSEM when performing comparative assessments between two building stock configurations, such as the greenhouse gas emissions case study presented at the end of journal article 3.



## REFERENCES

- Akshay Uttama Nambi, S. N., Antonio Reyes Lua, and R. Venkatesha Prasad. 2015. “LocED: Location-Aware Energy Disaggregation Framework.” In *BuildSys 2015 - 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*, 45–54. Association for Computing Machinery, Inc. doi:10.1145/2821650.2821659.
- Aksoezen, Mehmet, Magdalena Daniel, Uta Hassler, and Niklaus Kohler. 2015. “Building Age as an Indicator for Energy Consumption.” *Energy and Buildings* 87: 74–86. doi:10.1016/j.enbuild.2014.10.074.
- ASHRAE. “Building Energy Modeling Professional Certification.” *ASHRAE*. <https://www.ashrae.org/education-certification/certification/bemp-building-energy-modeling-professional-certification>.
- ASHRAE. 2013. *ASHRAE Handbook of Fundamentals*. Atlanta, GA: ASHRAE.
- Ballarini, Ilaria, Stefano Paolo Corgnati, and Vincenzo Corrado. 2014. “Use of Reference Buildings to Assess the Energy Saving Potentials of the Residential Building Stock: The Experience of TABULA Project.” *Energy Policy* 68: 273–284. doi:10.1016/j.enpol.2014.01.027.
- Barker, Sean, Aditya Mishra, David Irwin, Emmanuel Cecchet, Prashant Shenoy, and Jeannie Albrecht. 2012. “Smart\*: An Open Data Set and Tools for Enabling Research in Sustainable Homes.” In *Proc. SustKDD '12*, 1–6.
- Beckel, Christian, Leyna Sadamori, Thorsten Staake, and Silvia Santini. 2014. “Revealing Household Characteristics from Smart Meter Data.” *Energy* 78: 397–410.
- Booth, A.T., R. Choudhary, and D.J. Spiegelhalter. 2012. “Handling Uncertainty in Housing Stock Models.” *Building and Environment* 48 (February). Pergamon: 35–47. doi:10.1016/J.BUILDENV.2011.08.016.
- Box, George E P. 1976. “Science and Statistics.” *Journal of the American Statistical Association* 71 (356): 791–799.
- Caputo, Paola, Gaia Costa, and Simone Ferrari. 2013. “A Supporting Method for Defining Energy Strategies in the Building Sector at Urban Scale.” *Energy Policy* 55: 261–270.

doi:10.1016/j.enpol.2012.12.006.

- Carroll, Paula, Tadhg Murphy, Michael Hanley, Daniel Dempsey, and John Dunne. 2018. “Household Classification Using Smart Meter Data.” *Journal of Official Statistics* 34 (1): 1–25.
- CER. 2012. *CER Smart Metering Project - Electricity Customer Behaviour Trial, 2009-2010*. 1st ed. Irish Social Science Data Archive. SN: 0012-00.
- Cerezo, Carlos, Julia Sokol, Christoph Reinhart, and Adil Al-Mumin. 2015. “Three Methods for Characterizing Building Archetypes in Urban Energy Simulation. A Case Study in Kuwait City.” In *14th Conference of International Building Performance Simulation Association*, 2873–2880. Hyderabad, India.
- Chalmers, Carl, William Hurst, Michael Mackay, and Paul Fergus. 2019. “Identifying Behavioural Changes for Health Monitoring Applications Using the Advanced Metering Infrastructure.” *Behaviour & Information Technology* 38 (11). Taylor and Francis Ltd.: 1154–1166. doi:10.1080/0144929X.2019.1574900.
- Crawley, Drury B., Jon W. Hand, Michaël Kummert, and Brent T. Griffith. 2008. “Contrasting the Capabilities of Building Energy Performance Simulation Programs.” *Building and Environment* 43 (4): 661–673. doi:10.1016/j.buildenv.2006.10.027.
- Crawley, Drury B., Linda K. Lawrie, Frederick C. Winkelmann, W.F. Buhl, Y. Joe Huang, Curtis O. Pedersen, Richard K. Strand, et al. 2001. “EnergyPlus: Creating a New-Generation Building Energy Simulation Program.” *Energy and Buildings* 33 (4). Elsevier: 319–331. doi:10.1016/S0378-7788(00)00114-6.
- Dall’O’, Giuliano, Annalisa Galante, and Marco Torri. 2012. “A Methodology for the Energy Performance Classification of Residential Building Stock on an Urban Scale.” *Energy & Buildings* 48: 211–219. doi:10.1016/j.enbuild.2012.01.034.
- Dascalaki, Elena G., Kalliopi G. Droutsa, Constantinos A. Balaras, and Simon Kontoyiannidis. 2011. “Building Typologies as a Tool for Assessing the Energy Performance of Residential Buildings – A Case Study for the Hellenic Building Stock.” *Energy and Buildings* 43 (12):

- 3400–3409. doi:10.1016/j.enbuild.2011.09.002.
- Deb, Chirag, Mario Frei, Johannes Hofer, and Arno Schlueter. 2019. “Automated Load Disaggregation for Residences with Electrical Resistance Heating.” *Energy and Buildings* 182 (January). Elsevier Ltd: 61–74. doi:10.1016/j.enbuild.2018.10.011.
- “ESP-r: Building Performance Simulation Tool.” <http://www.esru.strath.ac.uk/Programs/ESP-r.htm>.
- Famuyibo, Adesoji Albert, Aidan Duffy, and Paul Strachan. 2012. “Developing Archetypes for Domestic Dwellings—An Irish Case Study.” *Energy and Buildings* 50: 150–157. doi:10.1016/j.enbuild.2012.03.033.
- Filogamo, Luana, Giorgia Peri, Gianfranco Rizzo, and Antonino Giaccone. 2014. “On the Classification of Large Residential Buildings Stocks by Sample Typologies for Energy Planning Purposes.” *Applied Energy* 135: 825–835. doi:10.1016/j.apenergy.2014.04.002.
- Firth, Steven K, and Kevin J Lomas. 2009. “Investigating CO2 Emission Reductions in Existing Urban Housing Using a Community Domestic Energy Model.” In *Building Simulation 2009*, 2098–2105. Glasgow, Scotland: Eleventh International IBPSA Conference.
- Fowler, K.M., A.H. Colotelo, J.L. Downs, K.D. Ham, J.W. Henderson, S.A. Montgmoery, S.A. Parker, and C.R. Vernon. 2015. *Simplified Processing Method for Meter Data Analysis*. Oak Ridge, TN.
- Fracastoro, Gian Vincenzo, and Matteo Serraino. 2010. “A Methodology for Assessing the Energy Performance of Large Scale Building Stocks and Possible Applications.” *Energy & Buildings* 43: 844–852. doi:10.1016/j.enbuild.2010.12.004.
- Gianniou, Panagiota, Christoph Reinhart, David Hsu, Alfred Heller, and Carsten Rode. 2018. “Estimation of Temperature Setpoints and Heat Transfer Coefficients among Residential Buildings in Denmark Based on Smart Meter Data.” *Building and Environment* 139 (July). Pergamon: 125–133. doi:10.1016/J.BUILDENV.2018.05.016.
- Government of Canada. 2021. “Engineering Climate Datasets.” *Engineering Climate Services*. [https://climate.weather.gc.ca/prods\\_servs/engineering\\_e.html](https://climate.weather.gc.ca/prods_servs/engineering_e.html).

- Government of Québec. 2020. *2030 Plan for a Green Economy*. Québec, Canada.
- Hammerstrom, D J, R Ambrosio, J Brous, T A Carlon, D P Chassin, J G Desteese, R T Guttromson, et al. 2007. *Pacific Northwest GridWise™ Testbed Demonstration Projects; Part I. Olympic Peninsula Project*.
- Hammerstrom, D J, J Brous, T A Carlon, D P Chassin, C Eustis, G R Horst, O M Järvegren, et al. 2007. *Pacific Northwest GridWise™ Testbed Demonstration Projects; Part II. Grid Friendly™ Appliance Project*.
- Harris, Chioke, Jared Langevin, Amir Roth, Patrick Phelan, Andrew Parker, Brian Ball, and Larry Brackney. 2016. “Scout: An Impact Analysis Tool for Building Energy-Efficiency Technologies.” In *2016 ACEEE Summer Study on Energy Efficiency in Buildings*, 4-1-4–12.
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. 12th print. Springer-Verlag.
- Heiple, Shem, and David J. Sailor. 2008. “Using Building Energy Simulation and Geospatial Modeling Techniques to Determine High Resolution Building Sector Energy Consumption Profiles.” *Energy and Buildings* 40 (8): 1426–1436.
- Higgins, J.A. 2012. “Energy Modeling Basics.” *ASHRAE Journal* 54 (12). American Society of Refrigerating Engineers: 26.
- Himeur, Yassine, Khalida Ghanem, Abdullah Alsalemi, Faycal Bensaali, and Abbes Amira. 2021. “Artificial Intelligence Based Anomaly Detection of Energy Consumption in Buildings: A Review, Current Trends and New Perspectives.” *Applied Energy*. Elsevier Ltd. doi:10.1016/j.apenergy.2021.116601.
- Hong, Tianzhen, Yixing Chen, Sang Hoon Lee, and Mary Ann Piette. 2016. “CityBES: A Web-Based Platform to Support City-Scale Building Energy Efficiency.” doi:10.1145/12345.67890.
- Hua, J., Z. Xiong, J. Lowey, E. Suh, and E. R. Dougherty. 2005. “Optimal Number of Features as a Function of Sample Size for Various Classification Rules.” *Bioinformatics* 21 (8). Oxford Academic: 1509–1515. doi:10.1093/bioinformatics/bti171.

- Hydro-Québec. 2012. *Réponse d'Hydro-Québec Distribution Aux Engagements 3, 10, 15, 18, 21 (UC), 21 (UMQ) ET 22 À 27.*
- Hydro-Québec. 2016. *Rapport Annuel 2015.* Montréal, Canada.
- Hydro-Québec. 2017. *Sustainable Development Plan 2020-2024.*
- Hydro-Québec. 2019a. “Pool Energy Calculator.”  
<http://www.hydroquebec.com/residential/customer-space/electricity-use/tools/swimming-pool-calculator.html>.
- Hydro-Québec. 2019b. “Spa Energy Calculator.”  
<http://www.hydroquebec.com/residential/customer-space/electricity-use/tools/spa-calculator.html>.
- Hydro-Québec. 2021. “Rate DT – Dual Energy | Hydro-Québec.”  
<https://www.hydroquebec.com/residential/customer-space/rates/rate-dt.html>.
- IEA. 2021. “IEA EBC Annex 70 – Building Energy Epidemiology.”  
<https://energyepidemiology.org/>.
- IEA (International Energy Agency). 2017. *Energy Efficiency 2017.*
- IEA (International Energy Agency). 2019. “Smart Grids - Tracking Clean Energy Progress.”  
<https://www.iea.org/tcep/energyintegration/smartgrids/>.
- James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning, with Applications in R.* Edited by Gareth M. James. 1st editio. Springer. doi:10.1007/978-1-4614-7138-7.
- James J. Hirsh and associates, and Lawrence Berkeley National Laboratory. 2021. “DOE-2 Building Energy Use and Cost Analysis Software.” <https://www.doe2.com/>.
- Johnson, Geoffrey, and Ian Beausoleil-Morrison. 2017. “Electrical-End-Use Data from 23 Houses Sampled Each Minute for Simulating Micro-Generation Systems.” *Applied Thermal Engineering* 114 (March). Pergamon: 1449–1456. doi:10.1016/J.APPLTHERMALENG.2016.07.133.

- Johnson, R. K. 2013. *Measured Performance of a Low Temperature Air Source Heat Pump Consortium for Advanced Residential Buildings*. Oak Ridge, TN, USA.
- Kasten, Fritz, and Gerhard Czeplak. 1980. "Solar and Terrestrial Radiation Dependent on the Amount and Type of Cloud." *Solar Energy* 24 (2). Pergamon: 177–189. doi:10.1016/0038-092X(80)90391-6.
- Kavgic, M., A. Mavrogianni, D. Mumovic, A. Summerfield, Z. Stevanovic, and M. Djurovic-Petrovic. 2010. "A Review of Bottom-up Building Stock Models for Energy Consumption in the Residential Sector." *Building and Environment* 45 (7): 1683–1697. doi:10.1016/j.buildenv.2010.01.021.
- Kleiminger, Wilhelm, Christian Beckel, and Silvia Santini. 2015. "Household Occupancy Monitoring Using Electricity Meters." In *2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2015)*, 975–986. doi:10.1145/2750858.2807538.
- Klein, S.A., W.A. Beckman, J.W. Mitchell, J.A. Duffie, N.A. Duffie, T.L. Freeman, J.C. Mitchell, et al. 2017. "TRNSYS 18: A Transient System Simulation Program, Solar Energy Laboratory, University of Wisconsin, Madison, USA." <http://sel.me.wisc.edu/trnsys>.
- Klemenjak, Christoph. 2018. "On Performance Evaluation and Machine Learning Approaches in Non-Intrusive Load Monitoring." *Energy Informatics* 1 (S1). Springer Science and Business Media LLC: 36. doi:10.1186/s42162-018-0051-1.
- Langevin, J., J. L. Reyna, S. Ebrahimigharehbaghi, N. Sandberg, P. Fennell, C. Nägeli, J. Laverge, et al. 2020. "Developing a Common Approach for Classifying Building Stock Energy Models." *Renewable and Sustainable Energy Reviews* 133 (November). Pergamon: 110276. doi:10.1016/J.RSER.2020.110276.
- Langevin, Jared, Chioke B. Harris, and Janet L. Reyna. 2019. "Assessing the Potential to Reduce U.S. Building CO<sub>2</sub> Emissions 80% by 2050." *Joule* 3 (10). Cell Press: 2403–2424. doi:10.1016/J.JOULE.2019.07.013.
- Loga, Tobias, Britta Stein, and Nikolaus Diefenbach. 2016. "TABULA Building Typologies in 20

- European Countries—Making Energy-Related Features of Residential Building Stocks Comparable.” *Energy and Buildings* 132: 4–12.
- Makonin, Stephen, Bradley Ellert, Ivan V. Bajić, and Fred Popowich. 2016. “Electricity, Water, and Natural Gas Consumption of a Residential House in Canada from 2012 to 2014.” *Scientific Data* 3 (160037): 1–12. doi:10.1038/sdata.2016.37.
- Marston, A., P. Garforth, G. Fleischhammer, and O. Baumann. 2014. “Urban Scale Modeling: How Generalized Models Can Help Communities Halve Their Energy Use in Thirty Years.” In *ASHRAE/IBPSA Conference 2014*. Atlanta, GA, USA.
- Mastrucci, Alessio, Olivier Baume, Francesca Stazi, and Ulrich Leopold. 2014. “Estimating Energy Savings for the Residential Building Stock of an Entire City: A GIS-Based Statistical Downscaling Approach Applied to Rotterdam” 75: 358–367. doi:10.1016/j.enbuild.2014.02.032.
- Mata, É., A. Sasic Kalagasidis, and F. Johnsson. 2014. “Building-Stock Aggregation through Archetype Buildings: France, Germany, Spain and the UK.” *Building and Environment* 81: 270–282. doi:10.1016/j.buildenv.2014.06.013.
- Mathworks Inc. 2018a. “Matlab R2018b.”
- Mathworks Inc. 2018b. “Matlab Statistics and Machine Learning Toolbox R2018b.” Natick, Massachusetts, United States.
- Mavrogianni, A, M Davies, M Kolokotroni, and I Hamilton. 2009. “A GIS-Based Bottom-up Space Heating Demand Model of the London Domestic Stock.” In *Building Simulation 2009*, 1061–1067. Glasgow, Scotland: Eleventh International IBPSA Conference.
- McKenna, Eoghan, and Murray Thomson. 2016. “High-Resolution Stochastic Integrated Thermal–Electrical Domestic Demand Model.” *Applied Energy* 165: 445–461.
- Miller, Clayton, Anjukan Kathirgamanathan, Bianca Picchetti, Pandarasamy Arjunan, June Young Park, Zoltan Nagy, Paul Raftery, Brodie W. Hobson, Zixiao Shi, and Forrest Meggers. 2020. “The Building Data Genome Project 2, Energy Meter Data from the ASHRAE Great Energy Predictor III Competition.” *Scientific Data* 7 (1). Nature Research: 1–13. doi:10.1038/s41597-

020-00712-x.

- Mordor Intelligence. 2021. “Global Smart Meters Market | Growth, Trends, Forecasts (2020 - 2025).” <https://www.mordorintelligence.com/industry-reports/global-smart-meters-market-industry>.
- Murray, David, Lina Stankovic, and Vladimir Stankovic. 2017. “An Electrical Load Measurements Dataset of United Kingdom Households from a Two-Year Longitudinal Study Background & Summary.” *Nature - Scientific Data*. doi:10.1038/sdata.2016.122.
- Nägeli, Claudio, Clara Camarasa, Martin Jakob, Giacomo Catenazzi, and York Ostermeyer. 2018. “Synthetic Building Stocks as a Way to Assess the Energy Demand and Greenhouse Gas Emissions of National Building Stocks.” *Energy and Buildings* 173 (August). Elsevier: 443–460. doi:10.1016/J.ENBUILD.2018.05.055.
- Najafi, Behzad, Monica Depalo, Fabio Rinaldi, and Reza Arghandeh. 2021. “Building Characterization through Smart Meter Data Analytics: Determination of the Most Influential Temporal and Importance-in-Prediction Based Features.” *Energy and Buildings* 234 (March). Elsevier Ltd: 110671. doi:10.1016/j.enbuild.2020.110671.
- National Renewable Energy Laboratory. 2021. “National Baseline Data Viewer - ResStock.” [https://resstock.nrel.gov/dataviewer/efs\\_v2\\_base](https://resstock.nrel.gov/dataviewer/efs_v2_base).
- Neale, Adam, Michaël Kummert, and Michel Bernier. 2018. “Generator : A Stochastic Virtual Smart Meter Data Generation Model for Residential Building Stock Characterization.” In *ESim 2018, the 10th Conference of IBPSA-Canada*, 65–74. Montreal, Canada.
- Neale, Adam, Michaël Kummert, and Michel Bernier. 2019. “Linear Discriminant Analysis for Classification of Building Parameters for a Large Virtual Smart Meter Data Set.” In *Proceedings of the 16th IBPSA Conference*, 3393–3400. Rome, Italy.
- Neale, Adam, Michaël Kummert, and Michel Bernier. 2020a. “Development of a Stochastic Virtual Smart Meter Data Set for a Residential Building Stock – Methodology and Sample Data.” *Journal of Building Performance Simulation* 13 (5): 583–605. doi:10.1080/19401493.2020.1800096.



- Neale, Adam, Michaël Kummert, and Michel Bernier. 2020b. “Virtual Smart Meter Data Set.” <https://vsmdata.meca.polymtl.ca/>.
- Neale, Adam, Michaël Kummert, and Michel Bernier. 2021. “Discriminant Analysis Classification of Residential Electricity Smart Meter Data.” *Energy and Buildings*, December. Elsevier, 111823. doi:10.1016/J.ENBUILD.2021.111823.
- New, Joshua Ryan, Mark B. Adams, Piljae Im, Hsiuhan Lexie Yang, Joshua C. Hambrick, William E. Copeland, Lilian B. Bruce, and James A. Ingraham. 2018. “Automatic Building Energy Model Creation (AutoBEM) for Urban-Scale Energy Modeling and Assessment of Value Propositions for Electric Utilities.” In *International Conference on Energy Engineering and Smart Grids*, 5. Oak Ridge, TN, USA: USDOE Office of Energy Efficiency and Renewable Energy, USDOE Office of Electricity.
- NRCan (Natural Resources Canada). 2011. *Survey of Household Energy Use*.
- NRCan (Natural Resources Canada). 2014. “Canada’s Secondary Energy Use by Sector, End Use and Subsector.” *Energy Use Data Handbook Tables*.
- NRCan (Natural Resources Canada). 2015. “2015 Survey of Household Energy Use Data Tables by Census Metropolitan Area.” *National Energy Use Database*. <https://oe.nrcan.gc.ca/corporate/statistics/neud/dpa/menus/sheu-cma/2015/tables.cfm>.
- NRCan (Natural Resources Canada). 2017a. “Energy Use in Canada: Trends Publications.” *National Energy Use Database*. <https://oe.nrcan.gc.ca/publications/statistics/trends/2017/totalsectors.cfm#L1>.
- NRCan (Natural Resources Canada). 2017b. “National Energy Use Database.” *Natural Resources Canada*.
- NRCan (Natural Resources Canada). 2018. “Energuide Housing Database.”
- NRCan (Natural Resources Canada). 2019a. “Residential Sector - Table 1: Secondary Energy Use and GHG Emissions by Energy Source.” *National Energy Use Database*. <http://oe.nrcan.gc.ca/corporate/statistics/neud/dpa/showTable.cfm?type=CP&sector=res&juris=qc&rn=1&page=0>.

- NRCan (Natural Resources Canada). 2019b. “Hot2000.” <https://www.nrcan.gc.ca/energy-efficiency/energy-efficiency-homes/professional-opportunities/tools-industry-professionals/20596>.
- NRCan (Natural Resources Canada). 2020. “Canada’s Secondary Energy Use (Final Demand) by Sector, End Use and Subsector.” <http://oee.nrcan.gc.ca/corporate/statistics/neud/dpa/showTable.cfm?type=HB&sector=aaa&juris=ca&rn=2&page=0>.
- NYSERDA. 2013. *Multifamily Performance Program Gas Furnace Electricity Usage Background*.
- Oak Ridge National Laboratories. 2021. “Virtual EPB.” [https://evenstar.ornl.gov/autobem/virtual\\_epb/](https://evenstar.ornl.gov/autobem/virtual_epb/).
- Oprea, Simona-Vasilica, Adela Bâra, Florina Camelia Puican, and Ioan Cosmin Radu. 2021. “Anomaly Detection with Machine Learning Algorithms and Big Data in Electricity Consumption.” *Sustainability 2021, Vol. 13, Page 10963* 13 (19). Multidisciplinary Digital Publishing Institute: 10963. doi:10.3390/SU131910963.
- Realtor.ca. 2019. “Comprehensive Real Estate Listing Database.” <https://www.realtor.ca/>.
- Régie de l’énergie. 2017. “Relevés Des Livraisons d’énergie En Vertu de l’entente Globale Cadre Pour La Période Du 1er Janvier Au 31 Décembre 2017 - Version Amendée.” *Audiences et Décisions, D-2016-143*. [http://www.regie-energie.qc.ca/audiences/Suivis/Suivi\\_HQD\\_D-2016-143.html](http://www.regie-energie.qc.ca/audiences/Suivis/Suivi_HQD_D-2016-143.html).
- Reinhart, Christoph F., and Carlos Cerezo Davila. 2016. “Urban Building Energy Modeling - A Review of a Nascent Field.” *Building and Environment* 97 (February). Elsevier Ltd: 196–202. doi:10.1016/j.buildenv.2015.12.001.
- Reinhart, Christoph F, Timur Dogan, Alstan Jakubiec, Tarek Rakha, and Andrew Sang. 2013. “UMI - an Urban Simulation Environment for Building Energy Use, Daylighting and Walkability.” In *13th Conference of International Building Performance Association*, 476–483. Chambéry, France.

- Saldanha, Neil, and Ian Beausoleil-Morrison. 2012. “Measured End-Use Electric Load Profiles for 12 Canadian Houses at High Temporal Resolution.” *Energy and Buildings* 49 (June). Elsevier: 519–530. doi:10.1016/J.ENBUILD.2012.02.050.
- Sherman, M.H., and D.T. Grimsrud. 1980. “Infiltration-Pressurization Correlation: Simplified Physical Modeling.” *ASHRAE Transactions* 86 (2): 778.
- Shimoda, Yoshiyuki, Takuro Fujii, Takao Morikawa, and Minoru Mizuno. 2004. “Residential End-Use Energy Simulation at City Scale.” *Building and Environment* 39: 959–967. doi:10.1016/j.buildenv.2004.01.020.
- Shmueli, Galit. 2010. “To Explain or to Predict?” *Statistical Science* 25 (3): 289–310. doi:10.1214/10-STS330.
- Sokol, Julia, Carlos Cerezo Davila, Christoph F. Reinhart, C. Cerezo, and Christoph F. Reinhart. 2016. “Validation of a Bayesian-Based Method for Defining Residential Archetypes in Urban Building Energy Models.” *Energy and Buildings* 134. Elsevier B.V.: 11–24.
- Statistics Canada. 2011. “Household Size, by Province and Territory.” *Census of Population and Statistics Canada Catalogue No. 98-313-XCB*.
- Statistics Canada. 2016. “2016 Census.” *Statistics Canada*. <http://www12.statcan.gc.ca/census-recensement/2016/dp-pd/index-eng.cfm>.
- Statistics Canada. 2019. “Annual Demographic Estimates: Subprovincial Areas.” <https://www150.statcan.gc.ca/n1/pub/91-214-x/2020001/section01-eng.htm>.
- Swan, Lukas G., and V. Ismet Ugursal. 2009. “Modeling of End-Use Energy Consumption in the Residential Sector: A Review of Modeling Techniques.” *Renewable and Sustainable Energy Reviews*. doi:10.1016/j.rser.2008.09.033.
- Swan, Lukas G., V. Ismet Ugursal, and Ian Beausoleil-Morrison. 2009. “A Database of House Descriptions Representative of the Canadian Housing Stock for Coupling to Building Energy Performance Simulation.” *Journal of Building Performance Simulation* 2 (2): 75–84. doi:10.1080/19401490802491827.

- Taniguchi-Matsuoka, Ayako, Yoshiyuki Shimoda, Minami Sugiyama, Yusuke Kurokawa, Haruka Matoba, Tomoya Yamasaki, Taro Morikuni, and Yohei Yamaguchi. 2020. “Evaluating Japan’s National Greenhouse Gas Reduction Policy Using a Bottom-up Residential End-Use Energy Simulation Model.” *Applied Energy* 279 (December). Elsevier: 115792. doi:10.1016/J.APENERGY.2020.115792.
- The Atmospheric Fund. 2019. *A Clearer View on Ontario’s Emissions Electricity Emissions Factors and Guidelines*.
- Theodoridou, Ifigeneia, Agis M Papadopoulos, and Manfred Hegger. 2011. “A Typological Classification of the Greek Residential Building Stock.” *Energy & Buildings* 43: 2779–2787. doi:10.1016/j.enbuild.2011.06.036.
- Transition Énergétique Québec. 2019. “Emission Factors and Conversion.” <https://transitionenergetique.gouv.qc.ca/fileadmin/medias/pdf/FacteursEmission.pdf>.
- TRNSYS. 2017. *Multizone Building Modeling with Type56 and TRNBuild*.
- Tuominen, Pekka, Riikka Holopainen, Lari Eskola, Juha Jokisalo, and Miimu Airaksinen. 2014. “Calculation Method and Tool for Assessing Energy Consumption in the Building Stock.” *Building and Environment* 75: 153–160. doi:10.1016/j.buildenv.2014.02.001.
- Ugursal, V I. 2017. *Canadian Building Stock Data: Notes from the Stock Modeling Workshop*. Ottawa, Canada.
- Ullah, Amin, Kilichbek Haydarov, Ijaz Ul Haq, Khan Muhammad, Seungmin Rho, Miyoung Lee, and Sung Wook Baik. 2020. “Deep Learning Assisted Buildings Energy Consumption Profiling Using Smart Meter Data.” *Sensors* 2020, Vol. 20, Page 873 20 (3). Multidisciplinary Digital Publishing Institute: 873. doi:10.3390/S20030873.
- US DOE. 2013. *EnergyPlus Engineering Reference - The Reference to EnergyPlus Calculations (v 8.1)*. Washington, DC, USA: US Department of Energy.
- Wang, Yi, Qixin Chen, Tao Hong, and Chongqing Kang. 2018. “Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges.” *IEEE Trans. Smart Grid*, 24. doi:10.1109/TSG.2018.2805.

- Westermann, Paul, Chirag Deb, Arno Schlueter, and Ralph Evins. 2020. "Unsupervised Learning of Energy Signatures to Identify the Heating System and Building Type Using Smart Meter Data." *Applied Energy* 264 (April). Elsevier: 14. doi:10.1016/J.APENERGY.2020.114715.
- Wilson, Eric, Craig Christensen, Scott Horowitz, Joseph Robertson, and Jeff Maguire. 2017. *Energy Efficiency Potential in the U.S. Single-Family Housing Stock*. Denver, CO, USA.
- Zhang, Leping, Lu Wan, Yong Xiao, Shuangquan Li, and Chengpeng Zhu. 2019. "Anomaly Detection Method of Smart Meters Data Based on GMM-LDA Clustering Feature Learning and PSO Support Vector Machine." In *ISPEC 2019 - 2019 IEEE Sustainable Power and Energy Conference: Grid Modernization for Energy Revolution, Proceedings*, 2407–2412. Institute of Electrical and Electronics Engineers Inc. doi:10.1109/iSPEC48194.2019.8974989.
- Zhang, Yang, Tao Huang, and Ettore Francesco Bompard. 2018. "Big Data Analytics in Smart Grids: A Review." *Energy Informatics* 1 (1). Springer Science and Business Media LLC: 8. doi:10.1186/s42162-018-0007-5.

## APPENDIX A SUMMARY OF BUILDING ARCHETYPE RESEARCH

A summary of building archetype research was presented by Reinhart & Cerezo Davila (Reinhart and Cerezo Davila 2016). Their work is adapted and expanded upon in Table A.1, with several archetype research works added. Table A.1 illustrates how different building stocks were segmented and the number of archetypes ultimately developed for each set of buildings. It should be noted that the table includes both residential and commercial cases.

It would appear that the number of archetypes per building for a building stock is not a linear relationship. The number of buildings (log scale) compared to the number of archetypes is illustrated in Figure A.1.

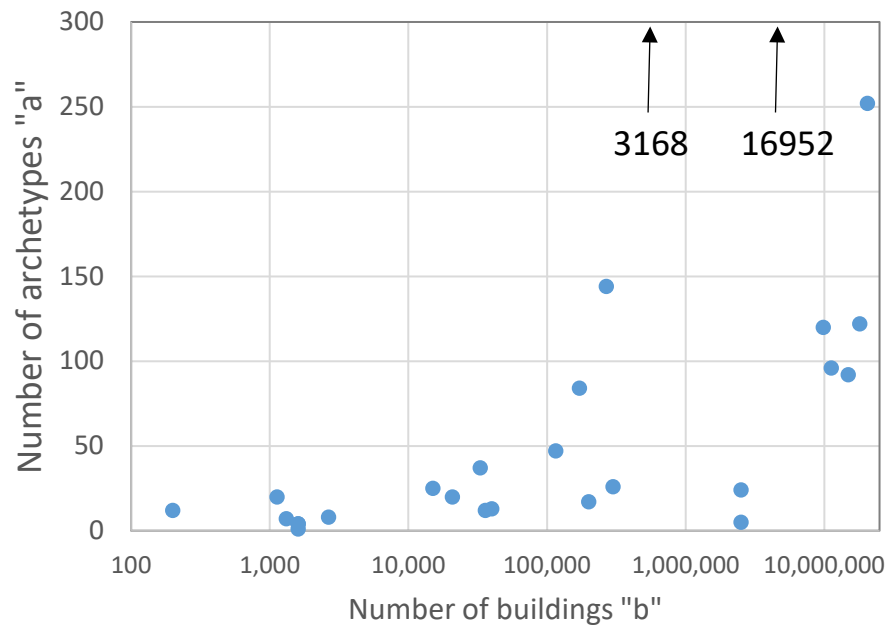


Figure A.1 Archetypes per number of buildings based on the literature

It is clear from Figure A.1 that there is a wide variety in the number of archetypes for each building stock studied. Many cases had 50 archetypes or less regardless of the number of buildings, so a general rule could indicate that most characterized stocks under 5 million buildings have been represented by 50 archetypes or less.

Table A.1 Review of building archetype research, adapted and updated from (Reinhart and Cerezo Davila 2016)

Application	# Buildings (b) <sup>i</sup>	Segmentation parameters	# Archetypes (a)	b/a	Characterization method <sup>iii</sup>	Ref.
Osaka (Urban)	1128	Shape, Area	20	56	Virtual	(Shimoda et al. 2004)
Houston (Urban)	<sup>ii</sup>	Shape, Age, Use, System	30		Virtual	(Heiple and Sailor 2008)
London (Urban)	267,000	Shape, Age	144	1854	Virtual	(Mavrogianni et al. 2009)
Carugate (Urban)	1320	Age	7	189	Sample	(Dall'O', Galante, and Torri 2012)
Milan (Urban)	<sup>ii</sup>	Shape, Age, Use	56		Virtual	(Caputo, Costa, and Ferrari 2013)
Rotterdam (Urban)	300,000	Shape, Age	26	11,538	Virtual	(Mastrucci et al. 2014)
USA locations (Urban)	200	Shape, Age, Use, System	12	17	Virtual	(Marston et al. 2014)
	33,000	Shape, Age, Use, System	37	892	Virtual	
	200,000	Shape, Age, Use, System	17	11,765	Virtual	
	15,000	Shape, Age, Use, System	25	600	Virtual	
Basel (Urban)	20,802	Shape, Age, Use	20	1040	Virtual	(Aksoezen et al. 2015)
UK (National)	115,751	Shape, Age	47	2463	Virtual	(Firth and Lomas 2009)
Italy (National)	11,226,595	Shape, Age, Climate	96	116,943	Sample	(Ballarini, Corgnati, and Corrado 2014)
Greece (National)	2,514,161	Shape, Age, Climate	24	104,716	Sample	(Dascalaki et al. 2011)
Greece (National)	2,514,161	Shape, Age, Use, System	5	502,832	Virtual	(Theodoridou, Papadopoulos, and Hegger 2011)
Italy (National)	11,226,595	Shape, Age, Climate, System	3168	277	Virtual	(Fracastoro and Serraino 2010)
Ireland (National)	40,000	Envelope, Thermal	13	3078	Virtual	(Famuyibo, Duffy, and Strachan 2012)
Sicily (Regional)	171,000	Shape, Age, Climate	84	2036	Virtual	(Filogamo et al. 2014)
France (National)	14,916,600	Shape, Age, Climate, System	92	162,137	Sample, Virtual	(Mata, Sasic Kalagasidis, and Johnsson 2014)
Spain (National)	9,804,090	Shape, Age, Climate, System	120	81,700	Sample, Virtual	
Germany (National)	18,040,000	Shape, Age, Climate, System	122	147,869	Sample, Virtual	
UK (National)	20,496,000	Shape, Age, Climate, System	252	81,333	Sample, Virtual	

Application	# Buildings (b) <sup>i</sup>	Segmentation parameters	# Archetypes (a)	b/a	Characterization method <sup>iii</sup>	Ref.
Finland (National)	36,000	Age, Use	12	3000	Sample	(Tuominen et al. 2014)
<b>Additional cases not covered by</b> (Reinhart and Cerezo Davila 2016)						
Canada (National)	7,191,540	Shape, Age, Climate, System, DHW	16,952	424	Sample	(Swan, Ugursal, and Beausoleil-Morrison 2009)
Cambridge (Urban)	2662	System, Envelope, Occupancy	8	333	Sample, Virtual	(Sokol et al. 2016)
Kuwait (Urban)	1600	N/A	1	1600	Virtual	(Cerezo et al. 2015)
Kuwait (Urban)	1600	Age	4	400	Virtual	
Kuwait (Urban)	1600	Age, probabilistic parameters	4	400	Sample, Virtual	

<sup>i</sup> Number of buildings represented by the archetypes

<sup>ii</sup> Data not available in the study

<sup>iii</sup> Virtual: based on modeling work, Sample: based on data set